

People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
University of Kasdi Merbah, Ouargla
Faculty of new information and communication technologies
Department of Computer Science and Information Technology



Thesis Submitted to the Department of Computer Science and Information
Technology in Candidacy for the Degree of “Doctor” 3rd Cycle LMD in
Computer Science

**Using machine learning techniques for automatic
annotation of personal image collections**

Option : Intelligence Artificielle et Technologies de l'Information

Presented by: Ramla Bensaci

Defended on June 29th, 2022
jury members:

President	Fatima Zohra Laallam	Pr. - Ouargla University
Examiner	Abd Elhakime Cheriet	MCA- Ouargla University
Examiner	Amine Khaldi	MCA- Ouargla University
Examiner	Mounir Beggas	MCA- El-Oued University
Examiner	Brahim Lejdel	Pr.- El-Oued University
Supervisor	Belal khaldi	MCA- Ouargla University

Academic -Year: 2021/2022

REMERCIEMENTS

We first thank the grace of Allah for having guided us and enlightened us on the right path of knowledge to continue this work and achieve the objectives traces.

I want to thank my family and friends for all the support throughout this work.

Last but not least, I would like to thank my supervisor, Dr. KHALDI Belal, for taking my thesis on for supervision and all the guidance and support throughout the work.

Ramla Bensaci

Dedications

I dedicate this work

To my dear Parents,

To my sister, my brothers, my nieces, and my nephews

To my husband Ayoub and my children Jasser, Tim Allah, and Rayan, who has helped me a lot with their patience and prayers

My dedications go tenderly to my dear friends Dr. Bouanane, Dr. Farah, Olaya,

To my colleagues in the imaging team,

To all my colleagues in the IT department.

To all who love me, and I love them

«Genius is 1% Inspiration and 99% Sweat(perspiration).»

Thomas Alva Edison (1847-1931)

«The value of an idea depends on its use.»

Thomas Alva Edison (1847-1931)

CONTENTS

Contents

List of Figures	I
List of Tables	III
Abstract	IV
Résumé	V
ملخص	VI
Chapter I : General introduction	1
1. General Introduction	2
1.1 Problems	4
1.2 Motivation	4
1.3 Contributions and Goals	5
1.4 Organization of the thesis	6
Chapter II: State of the art	8
1. Introduction	9
2. Image annotation Techniques	9
2.1 Manual annotation	9
2.2 Semi-automatic annotation	11
2.3 Automatic annotation	11
3. Automatic Image Annotation-model based	13
3.1 Generative models	13
3.2 Discriminative models	19
3.3 Nearest-Neighbor Model	20
3.4 Graph-based models	22
3.5 Deep learning model	26
4. Criteria for evaluating annotation systems	29
4.1 Per-label evaluation metrics	29
4.2 Per-image evaluation metrics	30
4.3 Per-annotation evaluation metrics	30
5. Image Databases	32
6. Conclusion	35
Chapter III: Feature Extraction and Segmentation	36
1. Introduction	37

CONTENTS

- 2. Feature Extraction 38
 - 2.1 Low-level feature-based AIA 38
 - 2.2 Domain-specific features:..... 47
- 3. Image segmentation..... 50
 - 3.1 Graph-based segmentation 51
 - 3.2 Contour-based segmentation..... 51
 - 3.3 Clustering-based segmentation..... 52
 - 3.4 Region-based segmentation..... 53
 - 3.5 Edge-based segmentation 54
 - 3.7 Grid-based segmentation 55
- 4. Conclusion 57
- Chapter IV: AIA Based on Machine Learning 58
 - 1. Introduction..... 59
 - 2. Unsupervised ML..... 60
 - 2.1 Clustering..... 60
 - 2.2 Hidden Markov Model (HMM) 60
 - 2.3 Artificial Neural Network..... 62
 - 3. Supervised-learning 64
 - 3.1 The k-Nearest Neighbor (k-NN)..... 65
 - 3.2 Decision Trees..... 65
 - 3.3 Support Vector Machine 66
 - 3.4 Naive Bayes 67
 - 4. Deep learning 68
 - 4.1 Deep Neural Network..... 68
 - 4.2 Deep convolutional neural networks 69
 - 5. Conclusion 74
- Chapter V: A framework for AIA 75
 - 1. Introduction..... 76
 - 2. Approach 77
 - 2.1 Image segmentation using algorithm JSEG 77
 - 2.2 Region representation..... 78
 - 2.3 Feature aggregation 79
 - 2.3 Calculating Blob/Label co-occurrences: 80
 - 2.4 Annotating new images..... 81

CONTENTS

3. Conclusion	82
Chapter VI: Experiments and result analysis.....	83
1. Introduction.....	84
2. Experiment Setup	84
2.1 Datasets.....	84
2.2 Evaluation Metrics:.....	85
2.3 Scenario 1: parameters tuning	86
2.4 Scenario 2: Comparing our method to the state of the art.....	89
2.5 Scenario 3: Computing cost.....	97
3. Conclusion	98
<i>Chapter VII: General conclusion.....</i>	<i>99</i>
1. conclusion.....	10100
2. Perspectives.....	10102
Bibliography.....	VII
Glossary	XXIX

List of Figures

Figure I.1 Generalized architecture for automatic image annotation and application in image retrieval on searching by the concept.....	Erreur ! Signet non défini.
Figure II.1. Maqam Echahid “The Martyrs’ Monument” Architecture in Algeria.	10
Figure II.2 : Examples of images from the test bases.....	33
Figure III.1. The basic scheme for image annotation as a classification problem	37
Figure III.2 .color models (RGB model - CMYK Color model - HSV model- LUV Color model - L*a*b model) respectively	40
Figure III.3 : illustration of 2D string: (a) an image decomposed into bloks, (b) object symbols are names, (c) definitions of relationship symbols, and (d) a 2D string	48
Figure III.4: Test results of different images segmentation.	50
Figure III.5: Various types of segmentation	51
Figure III.6 : : Result after use K-Means Clustering Algorithm (a), (d) Original image ; (b), (e) K-means algorithm ;	53
Figure III.7. Depicts a region segmentation result of an original image	54
Figure IV.1 : diagrammatic representation of ML techniques	59
Figure IV.2. A simple Neural Network.....	62
Figure IV.3. flowchart of the supervised learning process.....	64
Figure IV.4 : An example of K-nearest neighbour assignment with K = 1 (left) and K = 4 (right) (best viewed in colour).....	65
Figure IV.5. Design of Decision Tree (DT) classifier.	66
Figure IV.7. Example of typical convolutional neural network architecture	70
Figure IV.8. connection between neurons of convolutional layer (blue) and the input volume(red).....	70
Figure IV.9. Local connectivity of convolutional layer	70
Figure IV.10. Examples of pooling layers	70
Figure V.1 : Detailed architecture model of the system showing all components and the data flow in between, from the data sources to the annotated images. Black solid arrows correspond to training images whereas the blue ones correspond to test images.	76
Figure V.2 : a General scheme of Texture-enhanced JSEG (T-JSEG) segmentation method ...	78
Figure V.3 : MobileNet Architecture.	79
Figure VI.1: precisions/recalls yielded using the three aggregation methods : BoVF, VLAD and FV.....	86
Figure VI.2 The impact of changing the value of k of KNN regressor on (a) the precision and (b) recall of our proposed method.....	87
Figure VI.3: The impact of using different CNN models on our proposed method. The impact is measured in terms of precision, recall and complexity.....	88
Figure VI.4: statistical description of how our proposed method learns the meanings of concepts accurately and in a balanced manner. Precision and recall are denoted by the	

List of Figures

letters P and R, respectively, and the following number denotes the number of concepts utilized in the experiment. 93

Figure VI.5: A blob chart of F1 in terms of precision and recall. The experiments were conducted on MSRC dataset, with 22 concepts, between MBRM (S. L. Feng et al., 2004), CNN-THOP(J. Cao et al., 2020), SMK+GRM(Lu & Ip, 2009), CNN-AT(Le, 2016) and the Zhang et al. (W. Zhang et al., 2020) on one hand, and our proposed method on the other hand.... 94

Figure VI.6. precision heatmap generated from the precision per concept produced by each method. Lower precisions are indicated by darker cells, and vice versa. The methods involved in this experiment are MBRM, SSK-CBKP, CNN-AT,CNN-ECC, E2E-DCNN(2019),CNN-THOP, Zhang et al. and our method..... 95

List of Tables

Table II-1: compare annotation techniques in terms of the characteristics and requirements needed by humans and machines, and present some of the Advantages and Disadvantages of each technique.....	12
Table II-2. Advantages and disadvantages of topical models, Relevance models, and Mixture models.....	19
Table II-3. compare annotation model and present some of the Advantages and Disadvantages of each model	28
Table II-4: Large Scale Dataset	33
Table II-5: Performance comparison of various annotation methods on Corel-5K, ESP Game and IAPR TC-12 datasets using GoogLeNet features (Dutta, 2019).....	34
Table III-1. Different color descriptions are compared.....	43
Table III-2: Contrast of texture feature	46
Table III-3 : Comparison And Evaluation Of Segmentation Algorithms	56
Table IV-1 : A comparison between methods.....	72
Table VI-1 : specifications of the used two datasets, Corel 5k and MSRC v2.	85
Table VI-2 : A comparison between our method and other recent related works in terms of Precision (P), Recall(R), F1 and N+. The involved works adopt one of the scenarios : considering 260 concepts or 374 concepts, as shown in column (N ^o Cpt).....	89
Table VI-3.A list of images with their respective ground truths and given annotations. Concepts in bold indicate that they are parts of the ground truth.	96
Table VI-4.Time consumed, in seconds, for annotating one image with five concepts.	97

Abstract

As imaging equipment has advanced, the number of photosets has increased, making manual annotation impractical, necessitating the development of accurate and time-efficient image annotation systems. We consider the fundamental Computer Vision problem of image annotation, where an image must be automatically tagged with a set of discrete labels that best describe its semantics. As more digital images become available, image annotation can help automatically archival and retrieval of extensive image collections. Image annotation can assist in other visual learning tasks, such as image captioning, scene recognition, multi-object recognition, and image annotation at the heart of image understanding. Much literature on AIA has been proposed, primarily in probabilistic modelling, classification-based approaches, etc. This research explores image annotation approaches published in the last 20 years. In this thesis, we study the image annotation task from two aspects. First, The focus is mainly on machine learning models and AIA techniques based on the basic theory, feature extraction method, annotation accuracy, and datasets. Second, we attempt to address the annotation task by a CNN-kNN framework.

Furthermore, we present a hybrid approach that combines both advantages of CNN and the conventional concept-to-image assignment approaches. J-image segmentation (JSEG) is firstly used to segment the image into a set of homogeneous regions. A CNN is employed to produce a rich feature descriptor per area. Then, a vector of locally aggregated descriptors (VLAD) is applied to the extracted features to generate compact and unified descriptors. After that, the not too deep clustering (N2D clustering) algorithm is performed to define local manifolds constituting the feature space, and finally, the semantic relatedness is calculated for both image-concept and concept-concept using KNN regression to grasp better the meaning of concepts and how they relate. Through a comprehensive experimental evaluation, our method has indicated a superiority over a wide range of recent related works by yielding F1 scores of 58.89% and 80.24% with the datasets Corel 5k and MSRC v2, respectively. Additionally, it demonstrated a relatively high capacity for learning more concepts with higher accuracy, which results in N+ of 212 and 22 with the datasets Corel 5k and MSRC v2, respectively.

Keywords: Automatic image annotation; machine learning techniques; Image segmentation; features extraction, deep learning, CNN.

Résumé

Au fur et à mesure que l'équipement d'imagerie a progressé, le nombre de photosets a augmenté, rendant l'annotation manuelle peu pratique, nécessitant le développement de systèmes d'annotation d'images précis et rapides. Nous considérons le problème fondamental de vision par ordinateur de l'annotation d'images, où une image doit être automatiquement étiquetée avec un ensemble d'étiquettes discrètes qui décrivent le mieux sa sémantique. Au fur et à mesure que de plus en plus d'images numériques deviennent disponibles, l'annotation d'images peut aider à l'archivage et à la récupération automatiques de grandes collections d'images. Étant au cœur de la compréhension des images, l'annotation d'images peut également aider à d'autres tâches d'apprentissage visuel, telles que le sous-titrage d'images, la reconnaissance de scènes, la reconnaissance multi-objets, etc. De nombreuses publications sur l'AIA ont été proposées, principalement dans la modélisation probabiliste et la classification approches, etc. Cette recherche explore les approches d'annotation d'images publiées au cours des 20 dernières années. Dans cette thèse, nous étudions la tâche d'annotation d'images sous deux aspects. Premièrement, l'accent est mis principalement sur les modèles d'apprentissage automatique et les techniques AIA basées sur la théorie de base, la méthode d'extraction de caractéristiques, la précision des annotations et les ensembles de données. Deuxièmement, nous essayons d'aborder la tâche d'annotation par un cadre CNN-kNN. De plus, nous présentons une approche hybride qui combine à la fois les avantages de CNN et les approches conventionnelles d'attribution de concept à image. La segmentation d'image J (JSEG) est d'abord utilisée pour segmenter l'image en un ensemble de régions homogènes, puis un CNN est utilisé pour produire un descripteur de caractéristiques riche par zone, puis un vecteur de descripteurs localement agrégés (VLAD) est appliqué au caractéristiques extraites pour générer des descripteurs compacts et unifiés. Ensuite, l'algorithme de clustering pas trop profond (clustering N2D) est exécuté pour définir les variétés locales constituant l'espace des caractéristiques, et enfin, la relation sémantique est calculée à la fois pour l'image-concept et le concept-concept en utilisant la régression KNN pour mieux saisir la signification des concepts. et comment ils se rapportent. Grâce à une évaluation expérimentale complète, notre méthode a indiqué une supériorité sur un large éventail de travaux récents liés en produisant des scores F1 de 58,89 % et 80,24 % avec les ensembles de données Corel 5k et MSRC v2, respectivement. De plus, il a démontré une capacité relativement élevée à apprendre plus de concepts avec une plus grande précision, ce qui se traduit par un N+ de 212 et 22 avec les ensembles de données Corel 5k et MSRC v2, respectivement.

Titre : Utilisation de techniques d'apprentissage automatique pour l'annotation automatique de collections d'images personnelles

Mots-clés : Annotation automatique des images ; techniques d'apprentissage automatique; Segmentation d'images ; extraction de caractéristiques, Apprentissage en profondeur, CNN

ملخص

مع تقدم معدات التصوير ، زاد عدد مجموعات الصور ، مما يجعل التعليقات التوضيحية اليدوية غير عملية ، مما يستلزم تطوير أنظمة تعليقات توضيحية للصور دقيقة وفعالة من حيث الوقت. نحن نعتبر مشكلة رؤية الكمبيوتر الأساسية للتعليق التوضيحي للصور ، حيث يلزم وضع علامة تلقائيًا على الصورة بمجموعة من الملصقات المنفصلة التي تصف دلالاتها بشكل أفضل. مع توفر المزيد والمزيد من الصور الرقمية ، يمكن أن تساعد التعليقات التوضيحية للصور في الأرشيف التلقائي واسترجاع مجموعات الصور الكبيرة. نظرًا لكونه في قلب فهم الصور ، يمكن أن يساعد التعليق التوضيحي للصور أيضًا في مهام التعلم المرئي الأخرى ، مثل التعليق على الصورة ، والتعرف على المشهد ، والتعرف على الكائنات المتعددة ، وما إلى ذلك. تم اقتراح الكثير من الأدبيات حول AIA ، بشكل أساسي في النمذجة الاحتمالية والتصنيف القائم نهج ... إلخ. يستكشف هذا البحث مناهج التعليقات التوضيحية بالصور المنشورة في العشرين عامًا الماضية. في هذه الرسالة ، ندرس مهمة التعليق التوضيحي للصور من ناحيتين. أولاً ، ينصب التركيز بشكل أساسي على نماذج التعلم الآلي وتقنيات AIA القائمة على النظرية الأساسية وطريقة استخراج الميزات ودقة التعليقات التوضيحية ومجموعات البيانات. ثانيًا ، نحاول معالجة مهمة التعليق التوضيحي بواسطة إطار عمل CNN-kNN. علاوة على ذلك ، قدمنا نهجًا هجينًا يجمع بين مزايا CNN وأساليب تخصيص المفهوم إلى الصورة التقليدية. يتم استخدام تجزئة الصورة -J- SEG أولاً لتقسيم الصورة إلى مجموعة من المناطق المتجانسة ، ثم يتم استخدام CNN لإنتاج واصف ميزة غنية لكل منطقة ، ثم يتم تطبيق ناقل الواصفات المجمع محليًا (VLAD) على الميزات المستخرجة لإنشاء واصفات مضغوطة وموحدة. بعد ذلك ، يتم تنفيذ خوارزمية التجميع غير العميق (تجميع N2D) لتحديد المشعبات المحلية التي تشكل مساحة الميزة ، وأخيرًا ، يتم حساب الارتباط الدلالي لكل من الصورة - المفهوم والمفهوم - المفهوم باستخدام انحدار KNN لفهم معنى المفاهيم بشكل أفضل وكيف ترتبط. من خلال تقييم تجريبي شامل ، أشارت طريقتنا إلى التفوق على مجموعة واسعة من الأعمال ذات الصلة الحديثة من خلال تحقيق درجات F1 بنسبة 58.89٪ و 80.24٪ مع مجموعتي البيانات Corel 5k و MSRC v2 ، على التوالي. بالإضافة إلى ذلك ، فقد أظهر قدرة عالية نسبيًا على تعلم المزيد من المفاهيم بدقة أعلى ، مما أدى إلى N + من 212 و 22 مع مجموعتي البيانات Corel 5k و MSRC v2 ، على التوالي.

العنوان: استخدام تقنيات التعلم الآلي للتعليق التوضيحي التلقائي لمجموعات الصور الشخصية

الكلمات الرئيسية: شرح تلقائي للصور ؛ تقنيات التعلم الآلي. تقسيم الصورة؛ شرح منطقة بالصور. التعلم العميق. CNN.

Chapter I : General introduction

1. General Introduction	2
1.1 Problems	4
1.2 Motivation	4
1.3 Contributions and Goals	5
1.4 Organization of the thesis	6

1. General Introduction

Technological advancement is becoming increasingly straightforward for people to capture various locations and activities. There are thousands, if not millions, of personal images that are frequently stored without significant labelling. As a result, finding desired photographs became a tedious and time-consuming task. There is a need to develop an image content analysis and management system because the number of digital images is growing in both public and private picture collections. The digital image collection is helpful if a user can find images with some desired content. The content management of pictorial data is an organized way to store and search images from the database. The image labelling procedure (image annotation) entails giving a picture one or more labels (tags) describing its content. This procedure may be used for a variety of tasks, including automatic photo labelling on social media (J. Chen et al., 2021), automatic photo description for visually impaired persons (Stangl et al., 2020), and automatic text production from photographs (Ben et al., 2021), among others. Since it takes a lot of time and effort, manual image labelling (tagging) is inconvenient for small collections and impossible for huge ones. Automatic image annotation (AIA) was developed to address these issues, and it has since become a vibrant and essential academic topic. AIA models concepts using pre-annotated photo collections that are already accessible. Then, this learned model is applied to labelling unidentified images or completing partial labelled ones. The objective of the automated system is to use the features of the image contents to understand the idea. Due to this, understanding and analysing, image content is a tremendous challenge in computer vision with great significance for image data management and the advances in artificial intelligence. The main objective of this thesis is to address this issue by improving annotation performance through accurate tagging.

Automatic image annotation aims to attach keywords or tag labels to un-annotated images. Keywords are the description of the content or objects in images.

As defined by (Fu et al., 2012):

” Automatic image annotation is to automatically assign a collection of keywords from a given dictionary to a given image. In other words, the input is the target image, and the output is a set of keywords that describe this image in the best possible way. A computer can easily

calculate low-level features from color, texture and shape. Still, it cannot give a semantic interpretation of these features, unlike a person who can easily infer semantics from an image.”

As a result, Automatic image annotation (AIA) can be considered a multi-class object recognition issue, a difficult task that remains unsolved in computer vision. Automatic picture annotation and visual object categorization are two types of challenges that are highly comparable.

The first objective is to determine which keywords from a learned lexicon would best convey a new image, and the second task is to recognize the presence of particular items depicted in it. The words "describe" and "identify" in the preceding rough definitions show where similar algorithms differ. An automatic image annotation algorithm captures a more general concept of the described image. A picture of the ocean, coast, and seashore, for example, could be used to teach the term "beach." Image classification algorithms are commonly used to solve object categorization challenges. These algorithms implicitly do learn and differentiate between classes. These algorithms may easily classify a new image given a set of image representations and their associated class labels. As a result, that object category can annotate it. Figure I.1. shows an automatic image annotation system.

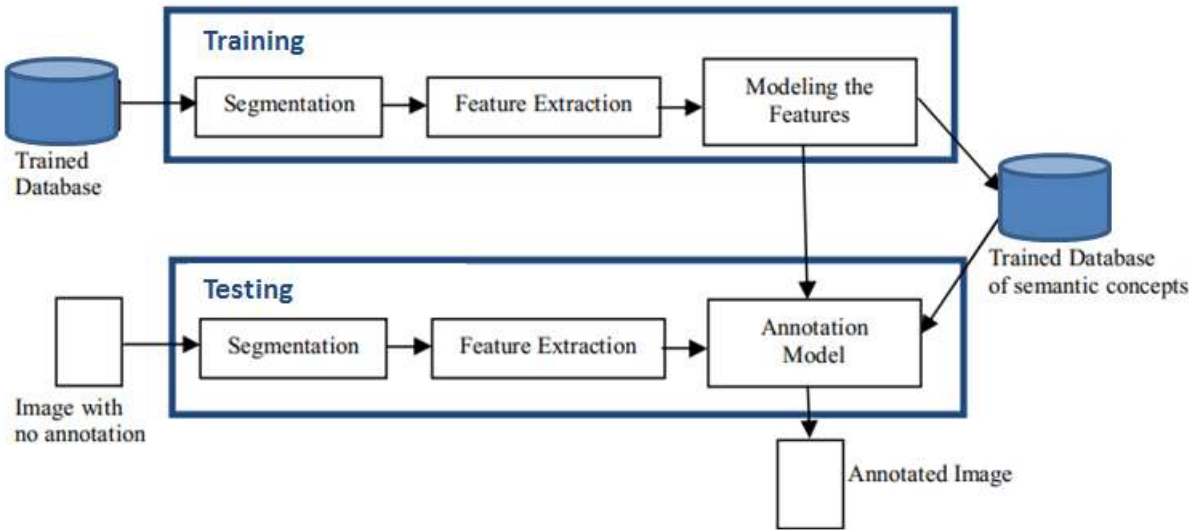


Figure I.1 Generalized architecture for automatic image annotation and application in image retrieval on searching by the concept

1.1 Problems

Automatic image annotation is a critical yet complex problem for computer vision researchers. Though automatic image annotation is a tough challenge for machine learning, intelligence can be applied to increase the performance of the annotation system by using machine learning. This problem can be solved by assigning semantic words to images. The ultimate goal is to use relevant machine learning techniques to create an effective automatic image annotation system. The system's speed and annotation accuracy are also being sought for improvement. This will improve content management performance and advance various applications, including Web image search, online photo sharing communities, and scientific experiments.

1.2 Motivation

The increasing need to manage substantial image sets is the primary motivation behind the research on image annotation. Due to a large amount of image data available on the internet, economical digital cameras, and increased storage capacity, there is an urgent need for an efficient annotation system.

Initially, Images are manually tagged with textual keywords in these approaches. Keywords can provide an exact representation of the semantics of images until the annotation is precise and comprehensive

Humans usually read the images semantically by combining the semantics of objects inside the images with the help of existing knowledge. The automatic image annotation system can do the same if intelligence is integrated. In particular, automatic image annotation plans to utilize existing annotated image datasets to tie visual features with keywords using machine learning techniques and predict the keyword for unannotated images. The results of state-of-art image annotation methods are unsatisfactory (C. Wang et al., 2007). Much work has been done on automatic image annotation to allow annotating images with minimal human assistance. The existing systems lack performance evaluation, and annotation's practical potential is mainly unaddressed (Huan Wang et al., 2008). Its performance is far from satisfactory due to the semantic gap between low-level visual features and high-level

symbolic concepts. Despite continued efforts in investigating new algorithms, developing a dedicated approach for annotation is advantageous.

1.3 Contributions and Goals

1.3.1 Goals

1-To organize the images semantically by tagging the images accurately with words to improve AIA.

2-Segmentation and feature extraction of images.

3-Model hybrid for image annotation by training.

4-Image annotation using machine learning.

5-Image Annotation for evaluation of annotation results.

1.3.2 Contributions

Though there are many methods available for segmentation, feature extraction, and auto-annotation using learning, designing an annotation system involves making many decisions, such as segmentation method, types of features used, extraction and representation of features, a machine learning method for annotation, combining of the methods and evaluation of the performance of the system. The contributions of the research work can be summarized as follows:

- ❖ An in-depth study into some of the quality issues with image data-sets used for image auto-annotation research.
- ❖ The thesis uses the generalized methodology of automatic image annotation, but an architectural design is proposed for auto-annotation. This proposed design is used during the implementation of the annotation framework:
 - Development of a segmentation system. Different segmentation methods are used on T-JSEG in Corel-5k and MSRC V2 datasets to segment the image into homogeneous regions.
 - Development of approaches to extract the low-level features of images and in-depth features. Geometric features from shapes, color, and texture extraction are carried out

to conclude the components selected for annotation. CNN is employed to produce a rich feature descriptor per region.

- Development of approaches to KNN regression has been employed to enhance both the representation of regions in the input feature space and the propagation of labels in the output semantic space
- ❖ we introduce a hybrid approach that combines the advantages of both CNN and the conventional concept-to-image assignment approaches
- ❖ Development of evaluation framework for assessment of annotation performance. The Corel image dataset's performance is explicitly improved for better classification categories like buildings and beaches.

1.4 Organization of the thesis

The structure of the thesis is organized as follows:

Chapter 2 – State of the art on Image annotation. This chapter describes prior and critical work done in the automated image annotation. This thesis uses techniques from these fields, and this chapter intends to allow the reader to place the above contributions in the context of previous work in these areas. Also, this chapter gives an overview of the significant components of an image annotation system.

Chapter 3 – Segmentation and Feature Extraction. The first step in the annotation system is to make regions of the image. In this chapter, segmentation techniques used in the system are presented. Segmentation of shapes using various edge detection techniques and color image segmentation using different segmentation techniques, along with evaluating these techniques, is presented. The second step in the annotation system is to extract the features representing images at a low level; their extraction and feature space representation are detailed. Shape, Color, and Texture feature extraction and replica are discussed. Finally, which features are advantageous, and why are they focused comparatively.

Chapter 4 – Automatic Image Annotation-based Machine Learning. This chapter deals with experiments and analysis of results for training and testing during annotation. The chapter is the heart of this thesis, where machine learning techniques used for annotation are

focused on their architectures. The experimental results and discussion on the classification-based performance of annotation are discussed.

Chapter 5 – Framework of the Annotation System. This chapter overviews the critical components of the automatic annotation system developed. The auto-annotation system developed uses the standard Corel and MSRC V2 image datasets. Details of these datasets are discussed in this chapter.

Chapter 6 – Experiments and result analysis: This chapter evaluates the retrieval that gives annotation evaluation.

Chapter 7 – Summary & Conclusions. In this chapter, the findings of the work are summarized, and conclusions of the overall result are discussed. Budding directions for future work are proposed to end up the discussion.

Chapter II: State of the art

1. Introduction	9
2. Image annotation Techniques	9
2.1 Manual annotation	9
2.2 Semi-automatic annotation	11
2.3 Automatic annotation	11
3. Automatic Image Annotation-model based	13
3.1 Generative models	13
3.2 Discriminative models	19
3.3 Nearest-Neighbor Model	20
3.4 Graph-based models	22
3.5 Deep learning model	26
4. Criteria for evaluating annotation systems	29
4.1 Per-label evaluation metrics	29
4.2 Per-image evaluation metrics	30
4.3 Per-annotation evaluation metrics	30
5. Image Databases	32
6. Conclusion	35

1. Introduction

This chapter gives an idea of the research related to the present thesis. The chapter reviews techniques and methods that serve as the basis for automatic image annotation. In the first part, we will look at a brief overview of manual, semi-automatic, and automatic annotation, emphasizing automatic annotation methods presented in the literature. The second part is devoted to presenting evaluation metrics for annotation systems and a description of the databases used in AIA. This is the research context for this thesis.

2. Image annotation Techniques

The annotation process involves assigning each image a keyword or set of keywords (or concepts) intended to describe the image's semantic content. These concepts can thus represent the low-level abstraction of an image (such as its color, its shape), the intermediate level (such as the objects contained in the image), or the high level (such as scenes and sensations). There are three types of annotation: manual annotation, semi-automatic annotation, and automatic annotation.

2.1 Manual annotation

The annotation process is done manually by users (human operators). Thus, an annotator assigns keywords (or concepts) to images based on his knowledge about the subject related to these images. Manual annotation of image content is considered the “best type” in terms of accuracy since keywords are selected based on the human determination of the semantic content of images. While manual annotation effectively provides a more appropriate description of the content of an image, this process is very time-consuming and expensive. We experimented and showed it to people to prove that the same idea can have several meanings. Three lay people were asked to describe the picture taken from real life (Figure II.1).

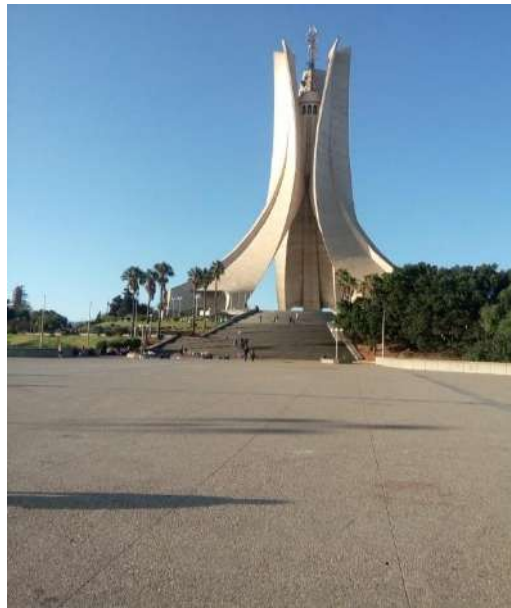


Figure II.1. Maqam Echahid “The Martyrs’ Monument” Architecture in Algeria.

- The first person is a 12-year-old child who described the contents of the picture: tall buildings, trees, Sky and Shade, etc.
- The second person is an amateur photographer: Maqam Echahid, Memorial, trees, shadows of people
- The third one is a mid-age regular person: Alger, Maqam Echahid, clear blue sky

This experience shows us that annotating an image varies from one person to another. It depends on his studies, culture, life, etc. Therefore, one of the disadvantages of the manual process is that it is subjective. Anyone can interpret (describe) an image according to their interests and personal point of view. In addition, this process is tedious, laborious, and time-consuming, especially for extensive collections. Moreover, the annotations provided can be inconsistent, general, ambiguous, noisy, incomplete, and sometimes inappropriate (Blei & Jordan, 2003; L. Wu et al., 2013), as nothing ensures seriousness during the whole process.

One solution to this problem is to annotate the same photo by multiple people and keep only shared annotations to get a 'public opinion' about the photo description. This solution requires multiple people. The development of the Internet and collaborative sites can also benefit people who use web images. There is no standard for the relationships between text and images included in web pages in web pages. Since web images are relatively broad in meaning, they were created by different groups for different reasons. Also, it is necessary to

create image bases used as basic facts. These are used to validate automatic annotation methods

2.2 Semi-automatic annotation

As the name suggests, semi-automatic annotation is divided into a first phase done manually and a second has done automatically(Suh & Bederson, 2004; Wenyin et al., 1999). The manual phase consists of choosing a representative sample of the collection of images and annotating it manually so that the annotations must be correct and complete (It requires the intervention of human annotators to generate initial descriptions of the images). This sample is generally referred to as a learning set. Then, the remaining images, called the test set, are annotated automatically using the training set in the second phase. Semi-automatic annotation is an intermediate solution between manual annotation and automatic annotation. It requires user intervention to annotate images or refine the automatic annotation results. Machine learning and user feedback help use previously annotated images to increase the annotation rate for images from the same domain. It is a consistent, cost-effective, fast, intelligent annotation of visual data. You can also take advantage of the Intelligent Image Indexing Web Service (Pagare & Shinde, 2012).

2.3 Automatic annotation

This is an entirely automatic process without any human intervention. Thus, the annotation system is responsible for extracting the characteristic concepts of a sample of images; in other words, the input is the target image. The output is a set of keywords that describe this image in the best possible way. The problem of automatic image annotation has been widely studied in recent years, and many approaches have been proposed to solve this problem.

Table II.1: compare annotation techniques in terms of the characteristics and requirements needed by humans and machines, and present some of the Advantages and Disadvantages of each technique

technique	Properties	Human and machine Requirements	Advantages	Disadvantages
Manual annotation	<ul style="list-style-type: none"> ✓ High accuracy in annotation ✓ There is subjectivity 	<ul style="list-style-type: none"> ✓ Needs a human expert to perform the annotation ✓ Save space to store and save data ✓ It takes a long time and has a high cost 	<ul style="list-style-type: none"> ✓ Provide complete descriptive information for retrieval or classification ✓ Accuracy in extracting semantic information on several levels. 	<ul style="list-style-type: none"> ✓ arduous requires a lot of time and effort, expensive ✓ There are many differences in the annotate according to the commentator
Semi-automatic annotation	<ul style="list-style-type: none"> ✓ The user can intervene in the form of relevant comments ✓ It can handle an incomplete data set 	<ul style="list-style-type: none"> ✓ Needs user intervention to improve feedback ✓ Using techniques to analyze and refine human descriptions ✓ It does not require a high-quality actual data set ✓ It requires the presence of the user during the annotation process 	<ul style="list-style-type: none"> ✓ Annotation efficiency ✓ More accurate and valuable for dynamic database 	<ul style="list-style-type: none"> ✓ Requires UI improvements to improve the feeding process
Automatic annotation	<ul style="list-style-type: none"> ✓ There is no subjectivity ✓ Consistency ✓ It can handle noisy, incomplete, and unstructured data sets 	<ul style="list-style-type: none"> ✓ It requires a precise training set ✓ Use identification techniques to create keywords ✓ It does not require the user to be present while annotation 	<ul style="list-style-type: none"> ✓ Saves time ✓ Accurate according to the data on which he is trained 	<ul style="list-style-type: none"> ✓ More prone to errors than manual comments ✓ Produces a more general (less detailed) annotation

3. Automatic Image Annotation-model based

A person can visually interpret the content of images and link topics with objects in the image (Z. Shi et al., 2017). Many researchers have sought to develop computer systems to simulate human ability in the early 1990s. Automatic image annotation (AIA) began to appear. The increasing emergence of digital images has played a significant and vital role in effective image retrieval, organization, classification, automatic annotation, etc. There has been extensive research on AIA and many methods and curricula. This section reviews the most used AIA approach in the past three decades, the model-based approach. This approach relies on training the annotation model from the features of the training set by connecting the visual attributes and textual metadata to annotate the unknown images. The model-based approach can be categorized into a generative model, discriminative model, graph-based model, nearest neighbour-based model, and deep learning-based model.

3.1 Generative models

A generative model is one of the first existing methods to learn a joint distribution over visual and contextual features so that the model can predict the conditional probability of tags given the image features. It is also customized to maximize the generative likelihood of image features and labels (Bhagat & Choudhary, 2018; Cheng et al., 2018). The generative models used for AIA consist of topical, relevance, and mixture models.

3.1.1 topical models

The topical model is one of the most widely used generative models for the AIA. Probabilistic topic models are a standard tool for measuring semantic meaning in inferred topics, and they work best on fixed probability, as they may conclude less linguistically significant topics that are unsupervised (J. Chang et al., 2009); The topic can be regarded as a means of image representation whose semantic level is higher than the visual feature. These models consider annotated images (training), samples from a specific mix of topics, each topic considering a probability distribution of image features and annotation words. Among the most famous works are :

in (2003) (Blei & Jordan, 2003) extended the LDA to correspondence LDA (cLDA) to learn to model the joint distribution of feature vectors associated with the regions of the image

and words of the caption, as well as to model the conditional correspondence between their respective reduced representations.

The probabilistic latent semantic analysis model (pLSA)(Monay & Gatica-Perez, 2004): assume that a group of co-occurrence words is associated with a latent topic. The topic is an intuitionistic concept characterized by a series of related words. For example, if “Apple” is regarded as a topic, then “ Steve Jobs” and “iPhone” probably appear frequently in this topic.

The work carried out by (C. Wang et al., 2008) uses Latent Dirichlet Allocation (LDA) to reduce the noise in the unbalanced labels of images and fully utilize the textual information for application in content-based images retrieval and search-based image annotation.

Later (Putthividhy et al., 2010) developed a novel image and video annotation model called topical regression multi-modal Latent Dirichlet Allocation (tr-mmLDA). Furthermore, tr-mmLDA can model correlations between data of different types. The correlations between the two data modalities were modelled using a linear Gaussian regression module, which allows a word to be connected with all image regions rather than just one. cLDA associates the word archery with a unique image region, and tr-mmLDA correlates it with all the image regions indistinguishably

In annotating satellite images (Bratasanu et al., 2011), LDA is also utilized to bridge the semantic gap between the outputs of automatic feature classification techniques and human-centred high-semantics.

Later (L. Song et al., 2016), the authors proposed sparse multi-modal topical coding (SMMTC), which is a new non-probabilistic formulation of the probability topical model (PTM) and extension of sparse topical coding (STC) for image annotation. SMMTC can capture more compact correlations between words and image regions.

Monay et al. proposed (Semantic et al., 2007) the famous PLSA-WORDS based on the Probabilistic Latent Semantic Analysis model (PLSA) to the co-occurrence of visual features and textual captions in annotated images. PLSA-WORDS uses aspects of one PLSA model to learn the semantic information from textual modality and then propagate it to the visual modality 2007

(D. Tian & Shi, 2020) proposed PLSA-MB, a two-stage hybrid probabilistic topic model to improve the quality of automatic image annotation by fusing PLSA with the max-bisection and then Integrating label and visual similarities of images associated with the labels

In 2020 (H. Song et al., 2020), a novel annotation approach based on the topic model, namely local learning-based PLSA (LL-PLSA), aims to improve the semantic level and reduce the complexity of model training(The LL-PLSA model learns from a local semantic neighbourhood consisting of only a tiny part of the training images) was proposed

Lienhart et al. (Lienhart et al., 2009) introduced a model called multilayer multimodal probabilistic Latent Semantic Analysis (mm-pLSA), which consists of two leaf-pLSAs (here derived from two separate input modalities: image tags and visual image features) and a single top-level pLSA node that merges the two leaf-pLSAs.

The MF-pLSA model (Rui Zhang et al., 2011) can be thought of as an extension of pLSA methods for image region annotation that combine low-level visual features in that it handles data from two different visual feature domains by adding one more leaf node to the graphical structure of the original pLSA

3.1.2 Relevance models

The relevance model-based AIA approaches to compute the posterior probability for each label of unlabeled images (typically the visual feature) by generating a combined distribution of image features (or regions) and tags (annotation keywords).

Finally, assign the tags for new photographs with the best chance of success. This is one of the most well-known works on the subject. Among them are the following:

Jeon et al. (Jeon et al., 2003) proposed the Cross-media Relevance Models (CMRM) to estimate the joint probability of visual and text-based semantic keywords. CMRM (Jeon et al., 2003) is subsequently improved by Continuous-space Relevance Model (CRM) (Lavrenko et al., 2004), which can directly model continuous features.

Wang et al. (C. Wang et al., 2007) proposed a content-based image annotation refinement (CIAR) algorithm, which formulates the image annotation refinement process as a Markov process and defines the candidate annotations as the states of a Markov chain.

In (S. L. Feng et al., 2004), the authors describe the Multiple-Bernoulli Relevance Model (MBRM), a statistical model for automatic annotating pictures and video frames. The Continuous-space Relevance Model (CRM) is the foundation of MBRM. The photos are divided into rectangles, and the features are extracted. They then learn a relevance model, a joint probability model for (continuous) image attributes and words, and annotate test photos. Multiple Bernoulli procedures are used to model words, while a kernel density estimate is used to model conditional random fields (CRF) images.

The sparse kernel relevance model (SKL-CRM) (Moran & Lavrenko, 2014b) introduces a sparse kernel learning framework into the continuous relevance model and greedily selects an optimal combination of kernels.

MRFs (Carbonetto et al., 2004) are typically formulated in a probabilistic generative framework modelling the joint probability of the image and its corresponding labels (Geman & Geman, 1984)

The CRF approach to the image labelling problem is broadened in this work (X. He et al., 2004), which is more complex due to the nature of 2D images against the one-dimensional nature of the text. It also seeks to learn the random field features that function at different image scales and then probabilistically combine the labelled images. Also, (Mensink et al., 2013) provide a tree-structured CRF model for interactive image labelling. Both employ CRFs to annotate multi-label images directly, but the structures of CRFs are different.

The authors (J. Zhang et al., 2015) proposed a new semantic-based image retrieval model in which images are segmented into semantic regions and then into grids. They then rely on CRFs for label correlations and ensure accurate automatic image annotation.

Wang et al. (Y. Wang et al., 2009) described a novel approach for automatic image annotation that uses an expanded cross-media relevance model to combine global, regional, and contextual information (CMRM). In a similar work (J. Liu et al., 2007), the authors used CMRM to study word-to-word relationships. The dual cross-media relevance model(DCMRM) combines word relation, image retrieval, and web search techniques to overcome the annotation challenge.

Additionally, Carneiro and Vasconcelos (Carneiro & Vasconcelos, 2005) segment each image into many blocks and describe each word class as a hierarchical mixture of Gaussians

describing the JPEG Discrete Cosine Transform (DCT) coefficient information of these blocks

In (Ben Rejeb et al., 2018), a fuzzy version of the Vector Approximation Files (VA-Files) was introduced, allowing for reliable multidimensional indexing to infer the associations between low-level features retrieved from region visual information and semantic notions. After getting the fuzzy codifications of regions, comparable clustering regions generate a joint distribution table to identify each cluster by distributing keywords that annotate its regions.

3.1.3 Mixture models

Mixture models formulate the image annotation problem to estimate the joint likelihood over visual features and words. To annotate an unseen test image, the model computes the conditional probability of each word in the vocabulary given the visual features of the image. A fixed number of the highest probability keywords are used as the annotation. Various mixture models have been developed for image annotation based on the parametric model. The basic idea is to learn the missing model parameters based on expectation-maximization methods (Dempster et al., 1977).

Wang et al. (C. Wang et al., 2009) utilized the universal Gaussian Mixture Models in a sparse coding framework for feature extraction and classification.

Another system uses automatic classification algorithms to extract and includes semantic information about the image content in the retrieval process. (Perronnin & Dance, 2007) presented a multi-level technique to annotate realistic situations' semantics by combining prominent visual components with relevant semantic ideas. To recognize the salient items automatically, Support Vector Machine classifiers with an automatic method for searching the optimal model parameters are used to learn a collection of detection functions from the annotated image regions. Finite mixture models approximate the class distributions of the relevant salient items to generate semantic notions.

In (Ruofei Zhang et al., 2006), a probabilistic semantic model was proposed, in which visual features and textual words are connected via a hidden layer, creating the semantic concepts to be discovered to harness the modalities' synergy explicitly. The association of visual elements and textual terms is decided in a Bayesian framework, allowing for confidence in the association.

Mori et al. (Yasuhide MORI et al., 1999) proposed a co-occurrence model to evaluate the correspondence between words and image regions using a uniform grid to predict annotated words for unseen images. However, this model requires many training samples to estimate a word probability.

Duygulu et al. (Duygulu et al., 2002) regarded the problem of AIA as analogous to machine translation in which one representation form (i.e., region) is desired to be translated to another (i.e., word). The 1:1 correspondence region/label can easily be modelled via a conventional EM algorithm by opting for such a model. After that, the authors presented two models for the joint distribution of text/blob and showed how image annotation is applied (Barnard et al., 2003).

The cross-media relevance model (Jeon et al., 2003) emerged and demonstrated the efficiency of learning the distribution of blobs and keywords. Blobs, in this context, result from clustering image features extracted from regions after using some typical segmentation algorithm. Instead of modelling blob–keyword via simple correlation, authors (S. L. Feng et al., 2004) modelled word probabilities using a multiple Bernoulli model and image feature probabilities using a nonparametric kernel density. In (B. Chen et al., 2020), the authors proposed a label co-occurrence learning framework based on graph convolution networks (GCNs) to directly examine the dependencies between pathologies for the multilabel chest X-ray. The works above require many training samples and have limited generalization ability to new categories.

Many AIA methods are inspired by generative models, which significantly contribute to AIA development. The generative models-based AIA approaches, on the other hand, have three major flaws. The first is generative models, which can estimate the generative likelihood of picture features and annotations but cannot guarantee tag prediction optimization. The second issue is that generative models may not capture the complex link between image attributes and labels. The third is the high computational demand imposed by complicated algorithms, e.g., the EM algorithm and the numerous parameter sets and parameter estimation procedures, which are typically computationally costly.

Table II.2. Advantages and disadvantages of topical models, Relevance models, and Mixture models.

Model	Advantages	Disadvantages
topical models	<ul style="list-style-type: none"> ✓ describe weakly annotated image content 	<ul style="list-style-type: none"> ✓ difficult to scale to a large-scale image dataset due to expensive storage cost or memory overhead
Relevance models	<ul style="list-style-type: none"> ✓ builds latent space in which the text and visual modalities are equally important. ✓ regional and contextual. features helps tag an image as a whole entity with semantic meaning. 	<ul style="list-style-type: none"> ✓ Weak relationships between visual feature co-occurrence and semantic content. ✓ all the regions within an image are assumed to be independently drawn from a generation probability distribution ✓ keyword propagation is only carried out from the training images to the test ones
Mixture models	-	<ul style="list-style-type: none"> ✓ the occurrence probability of a word is only related to one topic, while some words are common in specific combinations of topics ✓ the exact inference is intractable in these models, and to compute the posterior distribution

3.2 Discriminative models

Discriminative model-based AIA methods present the image annotation as a multi-label classification problem. To resolve this issue, Each label is considered a class, and binary classifiers are trained separately for each label using the visual features of the image. The trained classifier predicts whether the test image belongs to the class (with certain tags for that image).

Most of the discriminative models are based on a support vector machine (SVM) or its variants (Moran & Lavrenko, 2014a): In (Cusano et al., 2003), The annotation is performed by a classification system based on a multi-class Support Vector Machine for classifying images regions in one of seven classes. In (Mueen et al., 2008), automatic multilevel code generation is proposed for image classification and multilevel image annotation. Lindstaedt et al. (Lindstaedt et al., 2009) proposed automated image classification by offline supervised learning of concepts from visual folksonomies. (Tommasi et al., 2008) proposed a multi-cue approach to automatic medical image annotation based on the support vector machine algorithm. Gao and Fan (Fan et al., 2008) used Multiple kernels learning SVM not only to identify specific objects in an image, where some different kernels (color histogram kernel,

wavelet filter bank kernel, interest point matching kernel), but also to incorporate concept ontology to group similar items and label a theme for the image.

Afterward, in (Fakhari & Moghadam, 2013). The decision tree is enhanced to have a combination of classification and regression has been employed for multi-labelling image annotation in which concepts and their corresponding ranks will be stored in each DT leaf node instead of storing only a concept or a rank.

In (Goh et al., 2005), for multiclass annotation, the binary SVM (support vector classification) is used for semantic prediction to classify images into one of 116 concepts, and one class SVM (support vector regression) is used for the prediction of the confidence factor of the predicted semantic tags. The confidence factors of the same concept are added together. The concept with the maximum cumulative confidence is the final decision.

In (Grangier & Bengio, 2008), A model for retrieving images from text queries is introduced. They introduced a loss inspired by ranking SVM and formalized the notion of margin for retrieval problems.

3.3 Nearest-Neighbor Model

Nearest-neighbor (NN)-based models are the most important in AIA. They assume that visually similar images are more likely to propagate standard labels. They first identify visually similar neighbors from a set of training images for a test image. Using the distance metric to select similar neighbours, the test image tags are then derived based on the labels of the matching images.

Makadia et al. (Makadia et al., 2008) proposed the Joint Equal Contribution (JEC), one of the most classical nearest neighbor models. The nearest neighbor algorithm utilizes global low-level image features and assumes that the most similar neighbor shares more likely labels. The nearest neighbor of the query image is computed by combining the basic distance metrics, and then by ranking, the keywords are assigned using a greedy label transfer. Then it picks labels from additional neighbors and considers their frequencies and co-occurrence with the initially set labels for further assignment.

Guillaumin et al.(Guillaumin et al., 2009) introduced a nearest-neighbor method (TagProp), which is combined with metric learning by maximizing the log-likelihood of tag prediction in the training data

Later .Verma Y. et al. (Verma & Jawahar, 2012) proposed a two-pass k-nearest neighbor (2PKNN) algorithm for image annotation. It is a two-step variant of the classical k-nearest neighbor algorithm. The first step of 2PKNN uses “image-to-label” similarities, while the second step uses “image-to-image” similarities and combines the benefits of both of them. First, it obtains the neighborhood set of an image. Then, it uses the weighted similarity of the image to predict image labels. A distinguishing characteristic of this approach is that it can address the problem of sparse labels. And it has been updated by) Verma & Jawahar, 2017(so that they benchmark using new features extracted from a generic convolutional neural network model and those computed using modern encoding techniques.

In (Bakliwal & Jawahar, 2016), a learning-based image annotation model is proposed. They leverage the image-to-image and image-to-tag similarities to decide the best set of tags describing the semantics of an image.

In (F. Tian & Shen, 2015) created a novel search-based image annotation method by learning label set relevance, aiming to annotate large-scale image collections in the real environment.

NMF-KNN (Kalayeh et al., 2014)constructs a multi-view matrix containing different visual features and tags and then predicts tags by jointly factorizing multiple matrices.

In (Lin et al., 2012), a novel model using tag-related random search over range-constrained visual neighbors of the to-be-annotated image called TagSearcher was proposed, aiming to improve the performance of nearest neighbor model-based AIA methods

The authors (Mayhew et al., 2016) evaluated the effectiveness of two pre-trained CNN networks (AlexNet and VGG16) and 15 different manual features classifiers in nearest-neighbour-based label 2PKNN (Verma & Jawahar, 2012) and TagProp (Guillaumin et al., 2009). According to experimental results, the annotation performance achieved by employing features obtained from a deep CNN outperforms manual features.

In (Xia et al., 2016), the proposed method for automatic image annotation based on multi-feature fusion and multi-label learning algorithm combines various features using the feature fusion technique. It then uses multi-label KNN to annotate images automatically.

In (Z. Feng et al., 2013), an extension of kernel metric learning (KML) (David R. Hardoon, 2003) called robust kernel metric learning (RKML), which is a distance calculation technique based on regression, is used to find the visually similar neighbors of an image. The majority-based ranking is used to propagate the labels to a query image.

(Ballan et al., 2014) Proposed a learning procedure based on Kernel Canonical Correlation Analysis KCCA, which finds a mapping between visual and textual words by projecting them into a latent meaning space. The learned mapping is then used to annotate new images using advanced nearest-neighbor voting methods.

(Xu et al., 2013) proposed a novel label-specific prediction model that can precisely discriminate each label in each neighborhood. The weight of each label is determined by its specific distance value rather than the previous global distance value.

Lin Zijia et al. (Lin et al., 2015) proposed a new method called TagSearcher based on the conditional probability model. A constrained range is used instead of an exact, fixed number of visible neighbors. We are considering image-dependent weights of visual neighbors, tag-dependent trust degrees of visual neighbors, and votes for a candidate tag from visual neighbors.

In (Mensink et al., 2012), two types of classifiers are used with distance learning metrics, whereas LMNN (Mensink et al., 2012) is used with KNN and multiclass logistic regression-based distance metric learning is used with the nearest class mean (NCM).

3.4 Graph-based models

The basic idea behind the graph-based models is to design a graph from the visual and textual features. The correlation between visual and textual features can be represented in vertices and edges, explaining their dependency. The data points (visual features of images) and the labels can be described as separate subgraphs, and edges represent the correlation among subgraphs. The semantical correlation among labels can be represented using interconnected nodes, which helps multi-label image annotation. The graph-based models can

also be used to find the correlation among labels. In such a case, vertices represent labels, and edges represent correlation among labels.

Multiple features from distinct viewpoints are concatenated (Hu et al., 2017) and used to annotate the photos using a graph-based semi-supervised annotation model. The authors used a clustering technique to build prototypes in feature and concept spaces to deal with the vast storage space required for many photos. The optimal subset of features in both areas is then picked using a feature fusion method. The closest cluster to any test image is chosen in feature and concept spaces.

In (Hu et al., 2017), multiple features from different views are concatenated and used with a graph-based semi-supervised annotation model to annotate the images. The authors generated a prototype in feature and concept space using a clustering algorithm to deal with the ample storage space required for many images. Then, the best subset of features in both spaces is chosen using a feature fusion method. Its nearest cluster is selected in feature and concept spaces for any test image. All candidates are modelled as a bipartite graph. Then a reinforcement algorithm is performed on the bipartite graph to re-rank the candidates. Only the highest-ranking scores are reserved as the final annotations (Rui, 2007).

In their novel context-aware multi-label learning model (CMIML), Ding Xinmiao et al. (Ding et al., 2016) introduced a framework that enables multi-instance learning through the context and label context. The model consists of a graph-based instance context and a label context constructed using several latent conceptions.

(L. Feng & Bhanu, 2016) used co-occurrence patterns and random walks to rerank concepts created by generative and predictive models. Describes the multiple kernel learning (MKL) method for image annotation. Multiple kernel refinement (MKR) (Jiu & Sahbi, 2017) is employed, represented as a multi-layered mixture of nonlinear activation methods based on deep multi-layer networks. Every method comprises intermediate or elementary kernels that combine to form a positive semi-definite deep kernel. The different methods for learning network weights and plugging them into SVM for image annotation tasks are presented. The MIL approach for discriminative feature mapping was introduced in (R. Hong et al., 2014), which investigated both negative and positive correlations of concepts for the image annotation problem.

(C. Wang, 2006) proposed a novel method for automatically refining image annotations. The potential annotations are determined using a relevant model-based approach that incorporates visual information. The candidate annotations are then re-ranked, with only the best remaining final annotations. They reformulate the image annotation refinement process as a graph ranking issue and solve it with the Random Walk with Restart (RWR) algorithm to fully utilize the confidence levels of the candidate annotations and the corpus information.

(Lei et al., 2015) designed an image annotation framework via social diffusion analysis based on the common-interest model to analyze social diffusion records, the feature extraction from diffusion graphs and common interests, and the automatic annotation by the learning-to-rank method. With the assumption that the diffusion pattern of an image in social networks is highly related to the relevance between image annotations and user preferences.

In (J. Liu et al., 2009), a graph learning framework for image annotation was proposed. Image-based graph learning is performed to obtain the candidate annotations for each image. The authors proposed a new Nearest Span Series (NSC) method to create the image-based graph. Its edge weights are derived from sequential statistical information rather than traditional pairwise similarities. Also, word-based graph learning is developed to improve relationships between images and words to get the final annotations for each image to enrich the word-based graph representation.

In (Tang et al., 2011), a novel kNN-sparse graph-based semi-supervised learning approach for simultaneously harnessing the labelled and unlabeled data was proposed. The sparse graph constructed by datum-wise one-vs-kNN sparse reconstructions of all samples can remove most of the semantically-unrelated links among the data. Simultaneously, it applies the approximate k nearest neighbors to accelerate the sparse graph construction without losing effectiveness.

In (G. Chen et al., 2009), the authors suggested a new approach to construct a hypergraph to capture the correlations among different categories. Each vertex represents one training instance, and each hyperedge for one category contains all the instances belonging to the same category. Then, an improved SVM-like learning system incorporating the hypergraph regularization, called Rank-HLapSVM, is proposed to handle the multi-label classification problems.

(Hua Wang et al., 2011) suggested a novel Bi-relational Graph (BG) model. It has been applied to automatic image annotation and semantic image retrieval tasks, comprising the data graph and the label graph as subgraphs connected by an additional bipartite graph induced from label assignments. By considering each class and its labeled images as a semantic group, they perform random walks on the BG to produce group-to-vertex relevance, including class-to-image and class-to-class relevances. The former can be used to predict labels for unannotated images, while the latter are new class relationships, called Causal Relationships (CR), which are asymmetric.

Chen et al. proposed (X. Chen et al., 2010) a new large-scale graph-based multi-label propagation approach by minimizing the Kull back-Leibler divergence of the image-wise label confidence vector and its propagated version via the so-called hashing-based l_1 - graph, which is efficiently derived with Locality Sensitive Hashing approach followed by sparse l_1 -graph construction within the individual hashing buckets. Then, an efficient and convergence-provoking iterative procedure is presented for problem optimization.

In (Su & Xue, 2015), A method has been suggested that takes advantage of the nearest neighbor-based and the graph-based methods by exploiting the graph learning method to propagate the labels on the graph corresponding to the K nearest neighbors of a test image. To acquire more effective graph weights for computing scores for each label, besides the similarity of visual features, this method also considers the similarity of two label sets, computed based on the label correlation that captures the semantic information between two labels. It also combines the image-to-label distance with the graph learning-based score to compute the final decision value for labeling.

In (Z. Chen et al., 2020), a semantic-independent nearest-neighbor graph model is proposed based on semantic structure and graph learning. Specifically, graph learning is used to produce the pre-annotation of images based on label propagation of nearest-neighbor images, which can improve the accuracy of weak labels. Then, the semantic structure and the word graph are introduced to fine-tune the image annotation, reducing the redundancy of the predicted labels.

3.5 Deep learning model

Over the last decade, deep learning approaches have excelled at image processing. Visual attention has also proven effective, with deep neural networks being used in various NLP and computer vision methodologies. Several research (Guiding long-short term memory for image caption generation - On the origin of deep learning - An analysis of object appearance information and context-based classification- Deep learning-based feature representation for automated skin histopathological image annotation) have documented its use for image annotation. Despite the widespread use of deep learning-based technologies to improve the implementation of AIA frameworks, previous works on AIA have developed several deep learning procedures, such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), and Deep Neural Networks (DNN), etc.

The objective of the work (G. Qi et al., 2017) is to provide a model for identifying the functional relationships between text and image features so that translational and intramodal labels can be directly transferred to annotate images. They generate a new topic space into which the text and images are mapped to perform the label transfer procedure. The transfer function, which aligns diverse text and image spaces, is learned using both the occurrence and training sets. They also adhere to the notion of parsimony, encoding as few topics as feasible to ensure regular alignment between text and graphics. Intermodal label propagation can propagate labels from any labeled text corpus to any new image after training the transfer function.

The work (Johnson, 2015) has improved the multilabel image annotation by designing a typically model image metadata parametrically and using image metadata nonparametrically to generate neighborhoods of related images using Jaccard similarities uses a deep neural network to blend visual information from the image and its neighbors.

(B. Wu et al., 2018) proposed a diverse and distinct image annotation (D2IA). It leverages a generative adversarial network (GAN) model to train D2IA, generating a relevant and distinct tag subset. The tags are relevant to the image contents and semantically distinct, using sequential sampling to form a determinantal point process (DPP) model. Multiple such tag subsets covering diverse semantic aspects or semantic levels of the image contents are generated by randomly perturbing the DPP sampling process.

In (Vinyals et al., 2017), the authors showed a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and can be used to generate natural sentences describing an image. The model is trained to maximize the likelihood of the target description sentence given the training image.

In (R. Wang et al., 2017) constructed a large-scale image annotation model MVAIACNN based on a convolutional neural network, as they suggested a Multitask Voting (MV) method, which can improve the accuracy of original annotation to a certain extent, thereby enhancing the training effect of the model. the MV method can also achieve the adaptive label.

Y. Niu et al. proposed (Niu et al., 2019) a novel two-branch deep neural network architecture comprising a very deep main network branch and a companion feature fusion network branch designed to fuse the multi-scale features computed from the main branch. And introduced a label quantity prediction auxiliary task to the main label prediction task to explicitly estimate the optimal label number for a given image. This model extracted rich and discriminative features capable of representing a wide range of visual concepts and generalized Deep Transfer Networks for Knowledge Propagation in Heterogeneous Domains.

In (Gong et al., 2013) in their model, the highly expressive convolutional network features were taken hold of to solve the multi-label image annotation problem, as a network architecture similar to (Krizhevsky et al., 2017) was used, which contains several interconnected dense convolutional layers as the underlying architecture. As they used the top-k ranking loss, inspired by (Weston et al., 2010), for embedding to train the network

Murthy et al. (Murthy et al., 2015) proposed a CCA-KNN model based on the Canonical Correlation Analysis (CCA) framework. The new framework helps model both textual features (word embedding vectors) and visual features (CNN features) of the data

The effect of the depth of the convolutional network on its accuracy in a large-scale image recognition setting was also investigated. The paper (Le, 2016) addresses the depth and design of the ConvNet architecture and its fix. They also steadily increased the depth of the network by adding more convolutional layers, which is possible due to the use of tiny (3×3) convolutional filters on all layers.

Table II.3. compare annotation model and present some of the Advantages and Disadvantages of each model

Model	avantage	desavantage
Generative	<ul style="list-style-type: none"> ✓ Conditional probabilistic distribution ✓ well-formed theory ✓ alternative number of labels. 	<ul style="list-style-type: none"> ✓ Require prior image segmentation ✓ expensive training and computation ✓ sensitive to noisy data ✓ parametric ✓ might not be optimal
discriminative	<ul style="list-style-type: none"> ✓ ulti-label ✓ graph framework usual ✓ computation-efficien 	<ul style="list-style-type: none"> ✓ Sensitive to label-imbalance ✓ classes relevance ✓ parametric
Nearest-neighbor	<ul style="list-style-type: none"> ✓ Conceptually simple ✓ non-parametric ✓ do not require prior image segmentation ✓ large dataset 	<ul style="list-style-type: none"> ✓ Sensitive to small dataset ✓ distance metric learning ✓ sensitive to cluster result ✓ fixed number of labels.
Graph-based	<ul style="list-style-type: none"> ✓ More interpretable rules ✓ Attribute balance, ✓ no over-fitting, ✓ allow missing 	<ul style="list-style-type: none"> ✓ Over-fitting ✓ samples in memory ✓ Tall tree, need to test multiple attribute value
Deep learning	<ul style="list-style-type: none"> ✓ Deal with mass data ✓ learn very complicated relationships ✓ derive Robust features ✓ no manual selection is required ✓ Obtain sufficiently side information ✓ alternative number of labels 	<ul style="list-style-type: none"> ✓ Local optimum ✓ Vast training images ✓ Training process cannot be controlled.

4. Criteria for evaluating annotation systems

In the literature, several quality measures for image annotation systems are utilized. (BOUZAYANI, 2018; Dutta, 2019; Kwasnicka & Paradowski, 2006), They can be classified into per-label, per-annotation, and per-image evaluation metrics. Since their proposal, per-label measures have been widely utilized to evaluate image annotation models, while per-image metrics have been widely employed in recent research (Johnson, 2015). We'll go over these metrics.

4.1 Per-label evaluation metrics

4.1.1 Precision and recall

Recall and Precision: Allow images in the assessment dataset to be labeled with any keyword. Let B be the number of images with the label that have been annotated correctly, and let C be the number of images with the same label in the ground truth. The precision will be B/A , and the recall will be B/C . The recall metric assesses the ability to retrieve relevant information, whereas the precision metric reflects the ability to refuse unrelated information. The performance of AIA models is commonly assessed using a composite of recall and precision. However, assessing an AIA model's performance solely based on recall and precision is challenging because both metrics contradict each other. Even when the images are tagged with more or fewer labels on the ground. AIA approaches perform forced annotation of test images with k (generally 5) labels. As a result, even though the model predicts all ground truth labels, the recall and precision may be skewed (Bhagat & Choudhary, 2018).

$$P = \frac{1}{|S|} \sum_{s \in S} \frac{|\text{images annotated correctly with label } s|}{|\text{images annotated with label } s|} \quad (1)$$

$$R = \frac{1}{|S|} \sum_{s \in S} \frac{|\text{images annotated correctly with label } s|}{|\text{images having label } s \text{ in the ground truth}|} \quad (2)$$

4.1.2 F-measure

Because the performance of AIA models cannot be thoroughly evaluated using either recall or precision, the F1-score is engaged to alleviate this shortcoming:

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (3)$$

F1-score is also used to assess the robustness of AIA approaches, with a higher F1 score indicating a more robust model.

4.1.3 N plus (N+)

N plus measure is calculated when annotations for all the test-set are generated. This metric counts how many correctly assigned keywords W to at least one test image. It also shows how many good recall values there are for each keyword. The term $N+$ is used to represent the value of the measure. Recall greater than 0 value is integral and is defined over $N+ \in \{0, \dots, W\}$. High $N+$ values are common in high-performing AIA models.

4.2 Per-image evaluation metrics

Per-image evaluation measures have been implemented in AIA's modern business. Per-image precision, recall, and mAP are all taken into account here. Suppose that there are n_1 labels in a ground-truth, and the model predicts n_2 labels during testing, of which n_3 predictions are correct ($(n_3 \leq n_1$ and $n_3 \leq n_2)$). The precision and recall for this image will be n_3/n_2 and n_3/n_1 , respectively. These values are averaged over all photos in the test set (percentage %) to get an average.

4.3 Per-annotation evaluation metrics.

Annotation measurements focus on the result of frame-by-frame annotation. First, measurements are calculated after annotating each image. Then, the average measurements are calculated for all the images in the test set (BOUZAYANI, 2018).

4.3.1 Accuracy

Accuracy (Pan et al., 2004) is one of the essential quality indicators for auto-annotation algorithms. Accuracy is a common abbreviation for Accounting. It tells you how many words result in annotation correctly. If all words are successfully anticipated, the measure takes the value 1, and if none of the words is accurately predicted, the measure will be 0.

Equation 4 expresses Acc and defines it over $Acc_t \in \langle 0; 1 \rangle$. The arithmetic mean of all photos in the test set is called Average Acc.

$$Acc_t = \frac{c_t}{l_t} \quad (4)$$

$$\overline{Acc} = \frac{1}{T} \sum_{t=1}^T Acc_t \quad (5)$$

Where c_t is the number of successfully predicted words in the image t annotation, and l_t is the length of the generated annotation. It's worth mentioning that the highest achievable accuracy for annotations that are longer than planned is equal to $l_t^{expected} / l_t^{received}$

4.3.2 Normalized Score

The next metric is a normalized score (NS) (Barnard et al., 2003; Glotin & Tollari, 2005; Monay & Gatica-Perez, 2003). It's identical to Acc, except it also adds a penalty for any words that are misannotated. Equation 6 defines NS, which is defined over $NS_t \in \langle -1; 1 \rangle$.

$$NS_t = \frac{c_t}{l_t} - \frac{i_t}{W - l_t} = Acc_t - \frac{i_t}{W - l_t} = \textit{sensibility} + \textit{specificity} - 1 \quad (6)$$

Where W is the dictionary's size, N denotes the number of wrongly predicted terms. Equation 7 gives the average NS, determined by comprehensive annotations in the test set. Because it is challenging to interpret NS values for separate annotations universally, this average value is frequently reported in the literature.

$$\overline{NS} = \frac{1}{T} \sum_{t=1}^T NS_t \quad (7)$$

In this thesis, we have chosen Per-label evaluation metrics like most state-of-the-artwork. We used the four evaluation metrics: recall, precision, and F1- measure. Furthermore, $N +$.

5. Image Databases.

1. **Corel-5K** (Duygulu et al., 2002): contains 4,500 training and 499 testing images. Each image is annotated with five labels, with 3.4 labels per image on average. This is one of the oldest image annotation datasets and was considered the de facto benchmark for evaluation until recently. Since most of the recent image annotation techniques are based on deep neural networks and require extensive training data, there has been a decline in the usage of this dataset.

2. **ESP Game** (Ahn & Dabbish, 2004): This dataset contains 18,689 training and 2,081 testing images, with each image being annotated with up to 15 labels and 4.7 labels on average. It was formed using an online game where two mutually unknown players must assign labels to a given image and score points for every standard label. This way, several participants perform the manual annotation task, thus making this dataset quite challenging.

3. **IAPR TC-12** (M. Grubinger, P. D. Clough, H. M'uller, 2006): It contains 17,665 training and 1,962 testing images. Each image is annotated with up to 23 labels, with 5.7 labels per image on average. Each image is associated with a long description in multiple languages in this dataset. (Makadia et al., 2010) extracted nouns from the descriptions in the English language and treated them as annotations. Since then, it has been used extensively for evaluating image annotation methods.

4. **NUS-WIDE** (Chua et al., 2009): This is the largest publicly available image annotation dataset, containing 269,648 images downloaded from Flickr. The vocabulary contains 81 labels, with each image annotated with up to 3 labels. On average, there are 2.40 labels per image. Following the earlier papers (Bao et al., 2012; Gong et al., 2013; Ouni et al., 2021), we discard the images without labels. This leaves us with 209,347 images that we split into ~ 125K images for training and ~ 80K for testing by adopting the split initially provided by the authors of this dataset.

5. **MS-COCO** (Colleges et al., 2014): This is the second-largest popular image annotation dataset and is primarily used for object recognition in scene understanding. It contains 82,783 training images and 80 labels, with each image being annotated with 2.9 labels on average.

For this dataset, the ground truth of the test set is not publicly available. Hence, we consider the validation set containing 40,504 images as the test set in our experiments

Table II.4: Large Scale Dataset

Dataset	Corel 5K	ESP Game	IAPR TC-12	NUS-WIDE
No. of images	5000	20770	19627	269648 (209347 annotated)
No. of labels	260	268	291	81
Train images	4500	18689	17665	110K (not fixed)
Test images	500	2081	1962	4K (not fixed)
labels per image	3.4, 4, 5	4.7, 5, 15	5.7, 5, 23	2.4, 2
images per label	58.6, 22, 1004	326.7, 172, 4553	347.7, 153, 4999	5701.3, 1682
No. of labels (mean-freq)	195 (75.0%)	201 (75.0%)	217 (74.6%)	

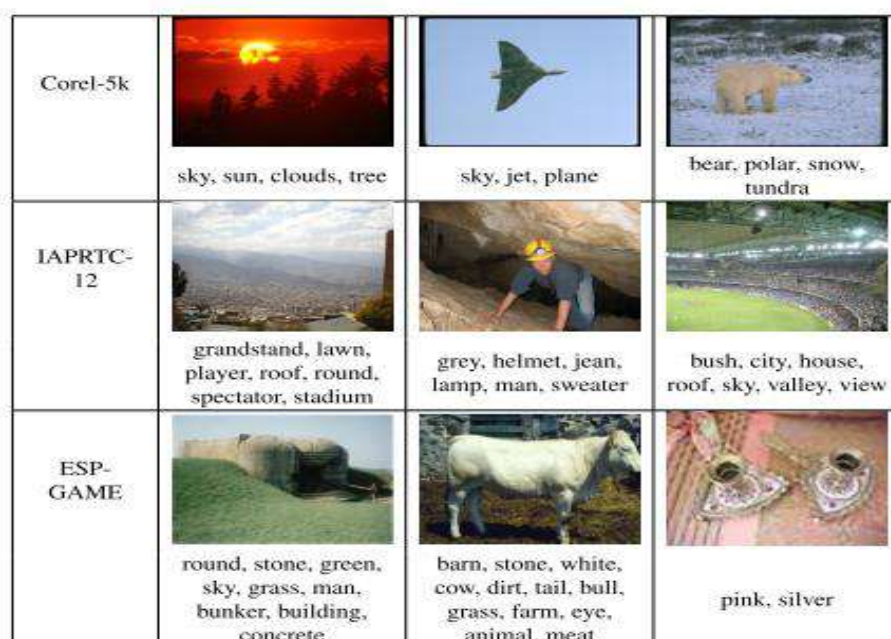


Figure II.2: Examples of images from the test bases

Table II 5: Performance comparison of various annotation methods on Corel-5K, ESP Game and IAPR TC-12 datasets using GoogLeNet features (Dutta, 2019)

Method	Per-label metrics				Per-image metrics		
	P	R	F1	N+	P	R	F1
Corel 5K							
JEC	41.70	44.95	43.27	161	45.97	64.92	53.76
TagPop	37.88	42.79	40.19	155	46.05	65.27	54.00
2PKNN	46.10	52.85	49.25	197	44.48	62.60	52.01
SVM	36	46.29	40.9	158	48.42	68.77	56.83
31.39IAPRTC 12							
JEC	44.52	27.77	34.20	226	49.62	47.92	48.76
TagPop	49.13	41.73	45.13	270	50.39	49.18	49.78
2PKNN	50.77	41.64	45.75	275	50.41	48.72	49.55
SVM	51.13	30.81	38.45	235	54.41	52.63	53.5
ESP Game							
JEC	45.15	31.39	37.03	239	41.85	47.06	44.31
TagPop	44.48	41.23	42.79	250	44.17	49.77	46.80
2PKNN	45.48	42.20	43.78	260	43.89	49.43	46.5
SVM	44.21	36.07	39.73	245	47.13	52.97	49.88

We conclude this chapter with a discussion of the different methods of AIA. Although manual image annotation is better for image retrieval because users choose keywords that describe the semantic content of images, it is labour-intensive and process. Tedious. Automatic annotation methods require a large number of examples to train. However, the annotated images available are not always sufficient in the real world. Therefore, a good compromise would be to choose an automatic annotation method and try to reduce the learning set. In this thesis, we place ourselves in the perspective of automatic annotation.

Now, using the annotation methods outlined in the section, we examine various elements of image annotation datasets and performance evaluation metrics (state of the art), Without forgetting that the databases have flaws, where Coral the annotations of this training set are not complete, has a shortage of And the **IAPRTC-12** has the problem of erroneous labels in the ground truth which negatively affect the performance of models proposed for AIA.

6. Conclusion

This chapter presents a detailed study of state of the art regarding image annotation. We discussed five types of AIA methods in terms of the ideas, models, algorithms, and open issues. We summarize the advantages and disadvantages of AIA methods in Table II.3. The generative model-based, the discriminative model-based, Graph-based models, and Deep learning model-based AIA methods are all learning-based methods. After this study, we presented the evaluation criteria for annotation systems. Then, we presented the databases used in this thesis. The conditional probability over images and labels is used to annotate images in generative model-based approaches. Image annotation is viewed as a multi-label classification problem by discriminative model-based AIA approaches. As a result, while the relationship between classes is essential, it cannot be answered directly by a binary classification system. For AIA, deep learning-based approaches typically use CNN to obtain robust visual features or alternative network frameworks, such as RNN, to exploit side information, such as label correlation.

In comparison, AIA methods based on the closest neighbor model use a two-step approach to annotate photos. Similar images are first fetched for the query image and then utilized to predict the question. The training process is not required for tag completion-based AIA approaches.

The next chapter will be devoted to describing our automatic image annotation model.

Chapter III: Feature Extraction and Segmentation

1. Introduction	37
2. Feature Extraction	38
2.1 Low-level feature-based AIA	38
2.2 Domain-specific features:	47
3. Image segmentation	50
3.1 Graph-based segmentation	51
3.2 Contour-based segmentation	51
3.3 Clustering-based segmentation	52
3.4 Region-based segmentation	53
3.5 Edge-based segmentation	54
3.7 Grid-based segmentation	55
4. Conclusion	57

1. Introduction

It is not suitable to put references and definitions in the introduction section

Image segmentation and feature selection are essential steps in automatic image annotation tasks. They provide data for the whole annotation process; the selected features strongly influence the annotation results: the better data produced, the better annotation results. AIA approach usually follows the scheme shown in Figure III.1. These, either extract :

- global features (computed on the whole image or by the use of dense sampling), or
- require prior segmentation of the image as regions/blobs or blocks.

Following that, these methods perform feature extraction. A feature is a way to capture a specific visual characteristic of an image, either globally or in a specific region. Color, texture, form, and conspicuous regions in images are the most often employed features for image annotation. A feature descriptor is used to apply a signature to the retrieved features. Finally, utilizing visual descriptors of images, a machine learning system is taught to recognize/detect concepts from the annotation vocabulary.

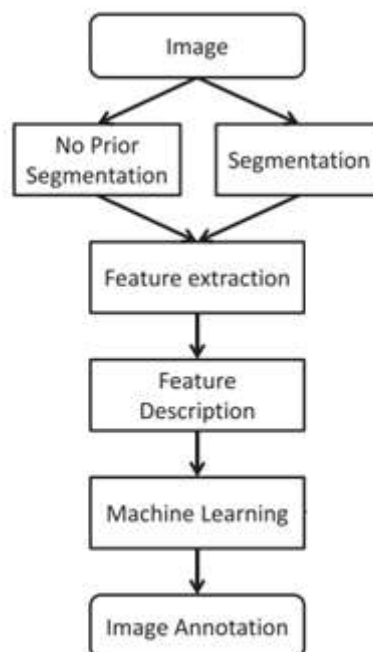


Figure III.1. The basic scheme for image annotation as a classification problem

2. Feature Extraction

Each image must be translated into a feature set that accurately depicts its visual contents for a learning algorithm to interpret the visual dataset. This takes the shape of one or more features. The extraction of features is the second phase in the annotation system. The core of picture understanding is feature extraction; these features can be classified as general or domain-specific. (Guang-Tsai et al., 1999)(D. Zhang et al., 2012).

General features: Color, shape, and texture are examples of application-independent features which can be further subdivided into the following categories (Monay & Gatica-Perez, 2004)(Carneiro et al., 2007):

- Pixel-level features characteristics computed at the pixel level, such as color and location.
- Local features: calculated features over the image's segmented areas or blocks due to subdivision.
- Global features: calculated features that span the entire image or a regular sub-area.

Domain-specific features: are a synthesis of low-level aspects for a specific domain, such as human faces, fingerprints, and conceptual features. On the other hand, all features can be divided into two categories: low-level features and high-level features. While low-level features can be extracted directly from the original images, high-level feature extraction requires low-level feature extraction.

2.1 Low-level feature-based AIA

This part discusses extracting prominent features with distinctive texture, shape, and color. The feasibility of extracting prominent visual features such as color, texture, and shape from images without user assistance is focused on.

2.1.1 Color features

Color is an essential feature for object recognition and matching images. In this section, color features are presented. Their invariance properties are summarized (F. et al., 2003)(D. et al., 1996)(S. et al., 2002)(G. R. C. & E., 2002)(Deng et al., 2001)(Borràs et al., 2003)(Kodituwakku & Selvarajah, 2004) Color features are widely used for image

representation because of their simplicity and effectiveness. Color features are extracted at both global and local levels of an image.

2.1.1.1 Color model:

In the color features, color models play a role; some popular color models used for automatic image annotation are presented below :

RGB model: the fundamental representation of color in computer. RGB uses an additive model in which red, green, and blue are combined in various ways to reproduce other colors. This color model is simple. It is also sensitive to illumination changes. This color model is widely used in object recognition and image annotation systems Blobworld (Carson et al., 1999).

CMYK Color model: The CMYK color model is a subset of the RGB model and is primarily used in color print production. CMYK is an acronym for cyan, magenta, yellow, and black (noted as K). The CMYK color space is subtractive, meaning that cyan, magenta, yellow, and black pigments or inks are applied to a white surface to subtract some color from the white surface to create the final color.

HSV model: Artists sometimes prefer to use the HSV color model over alternative models such as RGB or CMYK because of its similarities to how humans perceive color. HSV encapsulates information about a color in terms that are more familiar to humans. Using this color model in object representation has shown its efficiency and independence to illumination changes.

LUV Color model: The CIE LUV color model is considered a perceptually uniform color model. The lightness scale is replaced with a scale called L that is approximately uniformly spaced and more indicative of the actual visual differences. Chrominance components are U and V.

HMMD Color model: disclosed based upon hue, the shade, the tone, the tint, and the brightness of a color, and a color quantizing method using the hue max-min diff (HMMD) color space.

L*a*b model: The CIE 1976 L*a*b color model, defined by the International Commission on Illumination, is the complete color model used conventionally to describe all the colors visible to the human eye. The three parameters in the model represent the lightness of the

color (L), its position between magenta and green (a^*), and its position between yellow and blue (b^*).

Considering a three-dimensional color space (x, y, z), quantized on each component to a finite set of colors corresponding to the number of bins N_x, N_y, N_z , the color of the image I am the joint probability of the intensities of the three color channels.

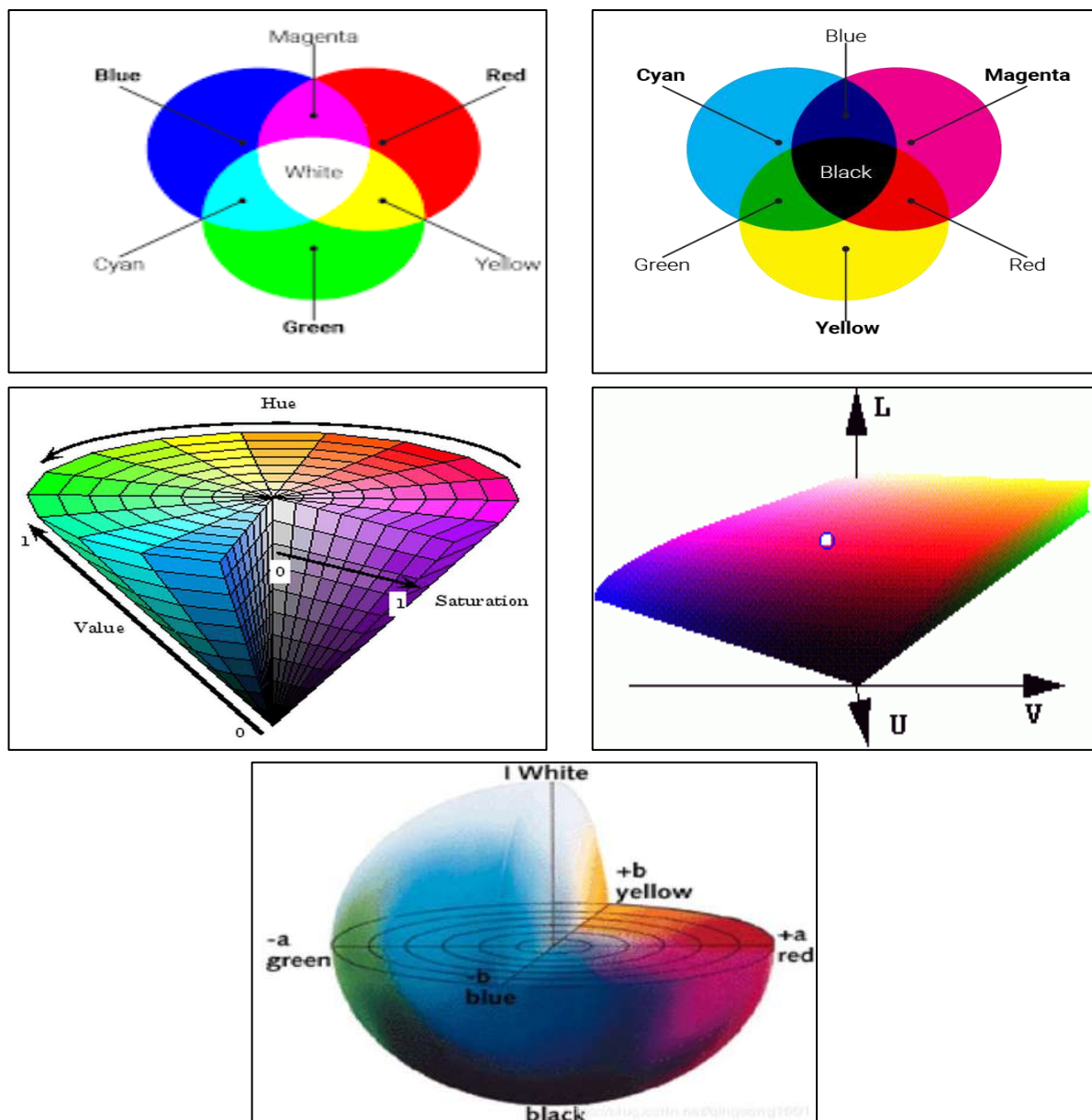


Figure III.2 .color models (RGB model - CMYK Color model - HSV model- LUV Color model - L^*a^*b model) respectively

Color is a crucial element for recognizing objects and matching images. Color features are described in this section and their invariance aspects. Because of their simplicity and efficacy,

color characteristics are commonly employed for image representation. Color characteristics are extracted at both the global and local levels of a picture. Color models play a role in color features. In figure II.2, we show the difference in Model Color types.

2.1.1.2 Color space:

Following the specification of the color space, color features are extracted from images or regions. In the literature, many important color characteristics have been offered, including color histogram (Berens et al., 2000), color moments (CM) (Flickner et al., 1995), color coherence vector (CCV) (Pass & Zabih, 1996), color correlogram (Huang et al., 1997), etc. MPEG-7 (S. et al., 2002) also standardizes many color features, including dominant color descriptor (DCD), color layout descriptor (CLD), color structure descriptor (CSD), and scalable color descriptor (SCD).

Colour moments: One of the most basic features is color moments. Many retrieval systems employ them (S. L. Feng et al., 2004) (Goh et al., 2005) (Flickner et al., 1995) (Fan et al., 2004) (Y. C. et al., 2005). The mean, standard deviation, and skewness are all frequent moments. They are usually calculated independently for each color channel (component). As a result, the feature vector is made up of nine features. When these features are determined for a region or item, they are valuable. However, the moments are sufficient to represent all of an image's color information.

Color histogram: The color histogram displays an image's color distribution (Goh et al., 2005) (Flickner et al., 1995) (Y. C. et al., 2005) (Vizza & Romani, 1809). It splits a color space into bins and counts the number of pixels corresponding to each color bin. Changes in translation and rotation do not affect this feature. Conversely, a color histogram does not reveal the spatial information of pixels. As a result, color histograms from visually different images can be comparable. Furthermore, a histogram's dimension is frequently quite large.

Color coherence vector: The color coherence vector (CCV) incorporates spatial information in the primary color histogram. Each histogram bin is divided into two sections: coherent and non-coherent parts. The pixels that are spatially related make up the coherent component. Isolated pixels are included in the non-coherent component. CCV usually outperforms a color histogram because it captures spatial information. A CCV, however, has twice the dimension of a traditional histogram.

Color correlogram: A color correlogram is a color version of the grey level co-occurrence matrix (GLCM). It describes how color pairs are distributed in an image (C. Wang, 2006)(C. Wang et al., 2006)(Huang et al., 1997). A color correlogram can be described as a three-dimensional histogram, with the first two dimensions representing the colors of any pixel pair and the third dimension representing the spatial distance between them (Huang et al., 1997). Each bin (i, j, k) in a correlogram represents the number of color pairs (i, j) at a distance of k . The horizontal distance $k=1$ is used to calculate the color correlogram. Correlograms for different distances can be calculated in the same way. Because it captures both intensity levels and spatial patterns in an image, the color correlogram outperforms the histogram and the CCV. Due to the large dimensionality and multiple matrix computation, it is much more complicated.

MPEG-7 color descriptors: The scalable color descriptor (SCD) is a histogram-based descriptor among MPEG-7 color descriptors. In HSV color space, SCD is essentially a histogram (X. Qi & Han, 2007). The scalability distinguishes it from the traditional histogram. Scalability is achieved in two ways: (1) by using the Haar transform to reduce the number of color bins, and (2) by deleting the least significant bits from the quantized (integer) representations of bin values. On the other hand, experimental data reveal that downscaling significantly impacts retrieval performance (S. et al., 2002). Furthermore, there is no spatial information in the description. As a result, it has a problem comparable to the traditional histogram.

Colour structure descriptor: A histogram-based descriptor is the Colour structure descriptor (CSD) (Kyung-Wook et al., 2007). A structuring element (such as a square) is moved throughout the image to constructing the CSD histogram. The histogram's bin I represent the number of times the structuring element has at least one pixel of color i . The CSD is an ordinary histogram if the window is 1 pixel in size. The performance of CSD is highly dependent on the window's size and structure, which are difficult to predict. It is also more computationally demanding than SCD.

Dominant color descriptor: A histogram variation is the dominant color descriptor (DCD)(Talib et al., 2013). DCD selects a small selection of colors from a histogram's highest bins. The bin height threshold determines the number of colors (bins) used as DCD. According to MPEG-7, 1–8 colors accurately depict a region. Unlike a standard histogram, DCD's selected colors adjust to the region rather than fixed in color space. As a result, DCD's

color representation is more accurate and compact than a traditional histogram. The calculation of similarity or distance between two DCDs, on the other hand, necessitates many-to-many matching.

Color moments are insufficient to express the regions among the numerous color attributes. On the other hand, histogram-based descriptors are either too high-dimensional or too time-consuming to compute. Color characteristics like CCV, color correlogram, and CSD help represent the entire image. However, they all need extensive computation. The varied color approaches are summarized in Table III.1.

Table III.1. Different color descriptions are compared.

Color method	pros	cons
Histogram	Simple to compute, intuitive	High dimension, no spatial info, sensitive to noise
CM	Compact, robust	Not enough to describe all colors, no spatial info
CCV	Spatial info	High dimension, high computation, cost
Correlogram	Spatial info	Very high computation cost, sensitive to noise, rotation, and scale
DCD	Compact, robust, perceptual meaning	Need post-processing for spatial info
CSD	Spatial info	Sensitive to noise, rotation, and scale
SCD	Compact on need, scalability	No spatial info, less accurate if compact

2.1.2. Texture features

While color is usually a pixel property, texture can only be calculated from a group of pixels, which is a well-researched picture feature. The texture feature is commonly employed in image retrieval and semantic learning due to its excellent discriminative capabilities. Multi-orientation filter banks (Malik & Perona, 1990) and the second-moment matrix (Frstner, 1994)(Gårding & Lindeberg, 1996) are two texture descriptors that have been proposed. We won't detail the traditional texture segmentation and classification approaches, which are both problematic and well-studied tasks. Instead, we use texture to add interest. They can be classified into two categories based on the extracted texture feature: spatial texture feature extraction methods and spectral texture feature extraction methods.

2.1.2.1 Spatial texture feature extraction methods

Texture features are extracted in the spatial technique by computing pixel statistics or locating local pixel structures in the original image domain. There are three types of spatial texture feature extraction techniques:

Structural: Texture primitives (exons or texture elements) and their placement criteria are used in structural techniques to describe textures (G. R. C. & E., 2002) (Nayak et al., 2017). The similarity of the two descriptors is determined using syntactical pattern recognition techniques.

Statistical: Texture measures low-level statistics of grey-level images by the statistical texture characteristic. Moments (G. R. C. & E., 2002)(F. et al., 2003), Tamura texture features (Islam et al., 2008)(Tamura et al., 1978)(Yavlinsky et al., 2005)(X. J. He et al., 2006), and features derived from the grey level co-occurrence matrix (GLCM) (F. et al., 2003)(Park et al., 2004) are all typical spatial domain statistical characteristics. Because statistical features are produced from a significant amount of data, they are compact and robust. They are, however, insufficient to depict the wide range of textures.

Model-based.: Texture is interpreted using stochastic (random) or generative models in model-based approaches. Model parameters describe the underlying textural property of the image. Markov random field (MRF) (F. et al., 2003)(Vailaya et al., 2001)(Nayak et al., 2017)(Cross, 1983)(Yavlinsky, 2007)(F. Liu & Picard, 1996)(Tuceryan & Jain, 1993)(Luo et al., n.d.), simultaneous autoregressive (SAR) model (J. Z. Wang et al., 2001), fractal dimension (FD) (Nayak et al., 2017)(Lions et al., 1995), and others are popular texture

models. These models are typically computationally demanding since they involve optimization.

2.1.2.2 Spectral texture feature extraction techniques

An image is transformed into the frequency domain, and then a feature is calculated from the transformed image in spectral texture feature extraction techniques. Fourier transform (FT) (K. Lee & Chen, 2005)(Hervé, 2007), discrete cosine transform (DCT) (Analysis, 2006), wavelet (Park et al., 2004)(Fan et al., 2004)(J. Z. Wang et al., 2001), and Gabor filters (S. et al., 2002)(Ruofei Zhang et al., 2006)(D. Zhang et al., n.d.) are all joint spectrum approaches. Although FT and DCT are quickly computed, they are not scaled or rotation invariant. Although Wavelet is efficient and reliable, it only collects horizontal and vertical features. Gabor features are the most resilient since they capture visual features in multiple orientations and scales. Curvelet features have recently been demonstrated to have considerable advantages over Gabor and wavelet features in multi-resolution analysis (Memon et al., 2017) since curvelet features are more effective in capturing curvilinear aspects, such as lines and edges(Starck et al., 2000).

(Islam et al., 2010) proposed a texture padding method to transform an unstable texture region into a square texture region. This method also acquires sizable regions to extract meaningful texture features

Both spatial and spectral features have advantages and disadvantages. Spatial features can be extracted from any shape without losing information and usually have semantic meaning understood- able by humans. However, acquiring many spatial features for image or region representation is challenging, and spatial features are usually sensitive to noise. On the other hand, spectral texture features are robust. They also take less computation because convolution in the spatial domain is done as a product in the frequency domain, implemented using FFT (B.S. Manjunath, 1996). However, they do not have the semantic meaning of spatial features. Spectral texture features are a desirable choice for images or regions with sufficient size. However, for small images or regions, especially when the regions are irregular, spatial features should be considered

Table III.2: Contrast of texture feature

	Pros	Cons
Spatial texture feature extraction methods	<ul style="list-style-type: none"> •Spatial texture methods are easy to understand, and many even have semantics. •They do not require regular region shapes and can be applied to irregular regions straightforwardly. •Spatial features can be extracted from any shape without losing information and usually have semantic meaning understood- able by humans 	<ul style="list-style-type: none"> • these features are usually sensitive to noise and distortions. • Many of these methods involve complex search and optimization processes with no general solutions.
Spectral texture feature extraction techniques	spectral texture features are robust take less computation because convolution in the spatial domain is done as a product in the frequency domain	<ul style="list-style-type: none"> • they can only be applied to square regions due to the use of FFT. • This method has the drawback that the blocks are too small to capture sufficient edge information

2.1.3 Shape features

The shape is essential for humans to identify and recognize real-world objects. Shape features have been employed for image retrieval in many applications. Zhang and Lu (D. Zhang & Lu, 2004) broadly classify shape extraction techniques into two significant groups: contour-based and region-based methods. Contour-based methods calculate shape features only from the shape's boundary, while region-based methods extract features from the entire region. Because contour-based techniques use only a portion of the region, they are more sensitive to noise than region-based techniques, as small changes in the shape significantly affect the shape contour. Therefore, color image retrieval usually employs region-based shape features.

Several simple region shape descriptors are commonly used in color image retrieval, including area, moments, circularity, and eccentricity. The area-based descriptor is used in several works (Yavlinsky, 2007)(Duygulu et al., 2002)(Y. C. et al., 2005)(Mezaris et al., 2003)(Jeon et al., 2003). Circularity and moments are used (Duygulu et al., 2002)(Y. C. et al., 2005)(Jeon et al., 2003). Circularity measures the ratio of area to the boundary. In (Mezaris et al., 2003), eccentricity or elongation is also used in the area. Eccentricity is the ratio of the

central axis's length to that of the minor axis. Individual simple shape descriptors are not robust. Therefore, they usually are combined to create a more helpful shape descriptor. More complex shape features are usually used in domain-specific applications such as trademark retrieval (Z. Hong & Jiang, 2008)(Avenue, n.d.) and object classification (Mezaris et al., 2003)(Kyung-Wook et al., 2007)(Y. Liu & Tjondronegoro, 2007), where the shape is an essential feature. For example, Park et al.(Kyung-Wook et al., 2007) use MPEG-7's contour shape descriptor, and Liu et al. (Y. Liu & Tjondronegoro, 2007) use the Fourier descriptor of shape contour for bird classification

2.2 Domain-specific features:

2.2.1 Spatial relationships

The spatial relationship tells object location within an image or the relationships between objects. The absolute spatial location of regions is used in (Fan et al., 2004)(Y. C. et al., 2005). Regarding the image, relative locations of regions, such as 'left, right, top, bottom, and centre' concerning the image, are used (Mezaris et al., 2003) for ontology-based concept learning. In (S.-K. Chang & Jungert, n.d.), the spatial relationship between regions is modelled using a 2D string. In a 2D string method, images are projected along the x- and y-axis. For each projection, an array of symbols represents the relationship between objects. The symbols are drawn from the set of object symbols and related symbols, such as 'left/right' and 'below/above'. Some variations of this method have been proposed (A. J. T. Lee & Chiu, 2003)(Y.I. Chang, B.Y. Yang, 2003). These approaches differ in the number of relational operators (symbols) and how they define those relations—the figure. III.3 shows an example of a 2D string representation. The image in Figure. III.3(a) is decomposed into regions (blocks). For simplicity, the block identifiers are used as object symbols. Two relationship symbols, 'o' and '¼' are used in this case. In horizontal and vertical directions, the symbol 'o' denotes 'left–right' and 'below–above' relationships. The symbol '¼' means the spatial relationship 'at the same spatial location as'. A 2D string takes the form (u, v), where u and v are the relationships of objects in horizontal and vertical directions, respectively. Figure. III.3(d) shows the 2D string for the figure image. III.3(a)

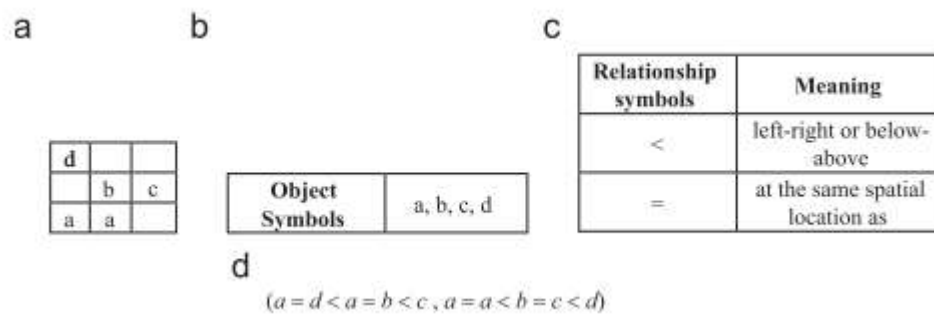


Figure III.3: illustration of 2D string: (a) an image decomposed into blocks, (b) object symbols are names, (c) definitions of relationship symbols, and (d) a 2D string

The 2D string and its variants can be used as global features for region-based representation, provided the segmented regions well define the objects. As segmentation algorithms often divide a single object into different fragments, the 2D string usually does not give an accurate representation. In practice, the relative location of regions is usually used (Mezaris et al., 2003; D. Zhang et al., 2009). In (D. Zhang et al., 2012), Zhang et al. define a distance relationship model for object location. It is assumed that different objects are located at different positions in an image. For instance, clouds and birds are usually at the top of an image, people and animals are usually at the centre, water and grass are usually at the bottom, etc. Therefore, objects can be differentiated based on the distance to their usual positions. Weight is defined based on the object's distance to its usual position. The weight is combined with other information such as color, texture, and shape to determine the object type.

2.2.2 Feature descriptors

To create salient features, salient properties in the image are utilized, commonly defined by color, texture, or local forms. Although color and texture qualities are frequently utilized to convey the contents of a picture, prominent spots provide additional distinguishing characteristics. Salient points operate as weak segmentation in the absence of object-level segmentation and play an everlasting role in the representation of an image. Significant points can appear throughout the image and do not have to be corners; they can even be smooth lines (Bhagat & Choudhary, 2018).

The conspicuous spots collected using wavelet and the corner detection approach were compared in ref (Sebe et al., 2003). Using local minima based on fractional Brownian (Pedersen, n.d.) is presented as a strategy for detecting conspicuous points and estimating

scale. Although salient points can be utilized in place of segmentation (since segmentation is a brittle process), they provide far more discriminative characteristics when used in conjunction with segmentation. Various authors have used salient features to annotate images (Fan et al., 2004).

- Scale-invariant feature transform (SIFT) (Lowe, 2004) based features have recently gained much traction. SIFT is a scale and rotation invariant local feature descriptor that collects vital points (interest points) and their descriptors from an edge-oriented histogram. SIFT uses a 128- dimensional descriptor.
- A resilient feature descriptor is the histogram of oriented gradients (HOG), which computes a histogram of the direction of gradients in a confined region of an image.
- Later (Bay et al., 2006), a speeded-up robust features (SURF) variant of SIFT was introduced. SURF has a 64-dimensional descriptor. SIFT and SURF descriptors are frequently translated into binary strings to speed up the matching process. SIFT and SURF are patented methods.
- Binary robust independent elementary features (BRIEF) provide a shortcut for directly locating binary strings without having to compute the descriptors. It's important to note that BRIEF is a feature descriptor rather than a feature detector.
- In (Rosten & Drummond, 2006) introduced a machine learning-based corner identification system called features from accelerated segment test (FAST) for real-time applications.

3. Image segmentation

Image segmentation breaks down the image f into connected regions $f_1, f_2, f_3, \dots, f_n$ according to a specific criterion of homogeneity (color, texture, etc.). The union of the regions should reproduce the initial image. Figure III.4 shows an example of the segmentation of images using the (Doggaz & Ferjani, 2011) algorithm.

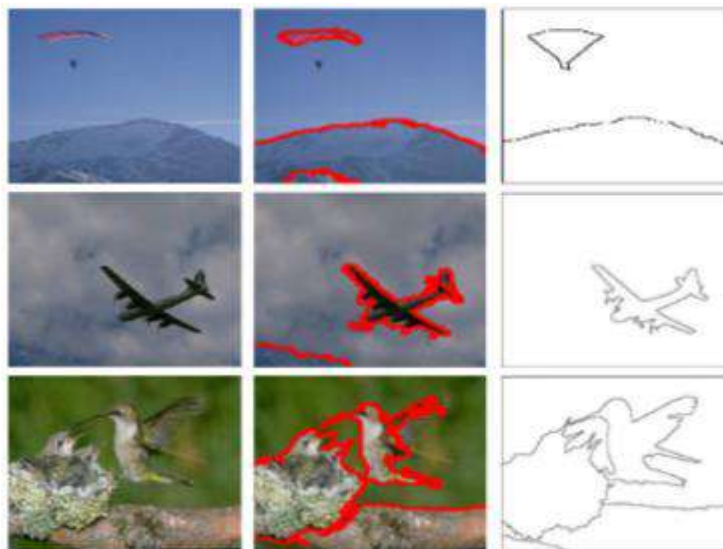


Figure III.4: Test results of different images segmentation

The regions produced might be regular or irregular. For example, the regular segmentation of the image results from splitting the image into square blocks of the same size (e.g. splitting the image into $32 \times 8 \times 8$ blocks). Irregular segmentation results from applying an algorithm such as N-cut or JSEG.

Image segmentation algorithms are generally classified into six categories (Bouchakwa et al., 2020)(Jaiswal & Pandey, 2021)(Yogamangalam & Karthikeyan, 2013), as specified in Figure III.5.

This section provides a brief review of segmentation methods commonly used in AIA.

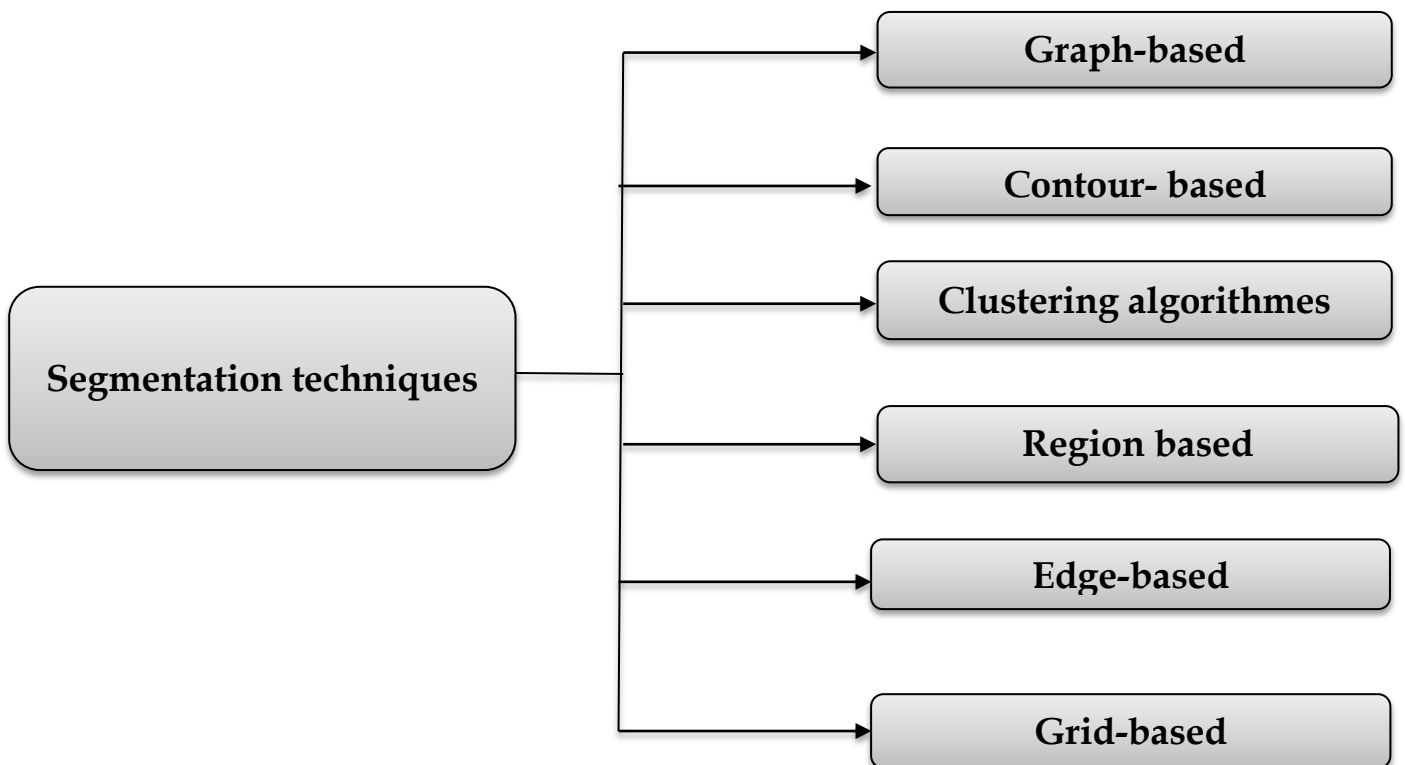


Figure III.5: Various types of segmentation

3.1 Graph-based segmentation

Image segmentation is modelled using graph-based methods by partitioning a graph into many sub-graphs, each representing a relevant object of interest in the image. The first step is mapping the image elements onto a graph $G = (V, E)$ where each node (also called vertices) $v_i \in V$ corresponds to a pixel in the image, and the edges in E connect specific pairs of neighboring pixels. Shi and Malik (J. Shi & Malik, 1997) propose a graph-based segmentation algorithm known as normalised cut (NCut). The NCut method represents an image as a graph where vertices are image pixels, and the edge weights represent the feature similarities between pixels. In (Duygulu et al., 2002) normalized cuts algorithm was used to create regions

3.2 Contour-based segmentation

The objective of contour-based segmentation is to create a curve around an item. The evolution comes to a halt when the curve meets the object's border. Contrary to cluster-based segmentation algorithms, contour-based segmentation algorithms do not require the number of clusters to be known beforehand (S. C. Zhu & Yuille, 1996). The underlying issue with this

approach is its reliance on precise edge detection, which is susceptible to noise. As a result, humans are frequently required to create a rough boundary outline, limiting the method's applicability to select domains, such as image processing tools.

3.3 Clustering-based segmentation

The purpose is to separate different homogeneous areas of an image and organize objects into groups (clusters) whose members have various properties in common (intensity, color, texture, etc.). We will limit ourselves to the study of the following segmentation methods:

3.3.2 K-Means Clustering Algorithm

The k-means algorithm (Dhanachandra et al., 2015; Likas et al., 2011) is the best-known and most widely used clustering algorithm due to its simplicity of implementation. It partitions the data of an image into K clusters. Unlike other so-called hierarchical methods, which create a "cluster tree" structure to describe clusters, k-means only creates a single level of clusters. The algorithm returns a data partition in which the objects inside each cluster are as close to each other and as far as possible from objects in other clusters. Each cluster in the partition is defined by its objects and its centroid. The k-means is an iterative algorithm that minimizes the sum of the distances between each object and the centroid of its cluster. The initial position of the centroids determines the final result, so the centroids should be initially placed as far apart as possible to optimize the algorithm. K-means changes cluster objects until the sum can no longer decrease. The result is a set of compact and separated clusters, provided that the correct K-value for the number of clusters has been chosen.

Let us consider an image with a resolution of $x \times y$, and the image has to be clustered into k number of clusters. Let $p(x, y)$ be an input pixel to be cluster and c_k be the cluster centres. The main steps of the k-means algorithm are following as:

1. Random choice of the initial position of the K clusters.
2. (Re-) Assign the objects to a cluster according to a distance minimization criterion d (generally according to a Euclidean distance measure).

$$d = \|p(x, y) - C_k\| \tag{21}$$

3. Once all the objects have been placed, recalculate the K centroids.

$$C_k = \frac{1}{k} \sum_{y \in C_k} \sum_{x \in C_k} p(x, y) \quad (22)$$

4. Repeat steps 2 and 3 until no more reassignments are made.
5. Reshape the cluster pixels into an image.

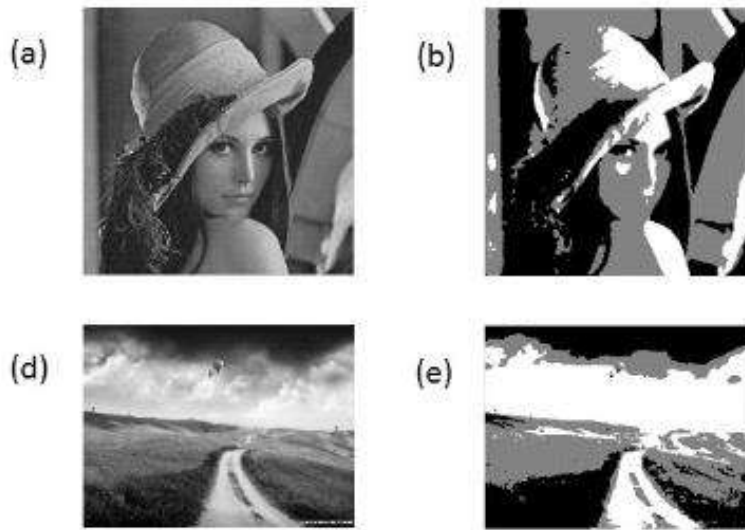


Figure III.6: Result after using K-Means Clustering Algorithm (a), (d) Original image ; (b), (e) K-means algorithm ;

3.4 Region-based segmentation

A region is a group of connected pixels with similar properties. It is essential in interpreting an image because it may correspond to objects in a scene. In the region-based segmentation, pixels corresponding to an object are grouped and marked. The partition into regions is done often by using values of the image pixels (Kaganami & Beiji, 2009; LALAOUI & MOHAMADI, 2013; Slabaugh et al., 2009).

For example, given a set of image pixels I and a homogeneity predicate $H(\cdot)$, let us partition the image I into a set of n regions R_i ; if equation (23) holds, then all pixels of any given region satisfy the homogeneity predicate $H(2)$. Also, two adjacent regions cannot be merged into a single region (25).

$$\bigcup_{i=1}^n R_i = True \quad (23)$$

$$\forall I, H(R_i) = True \quad (24)$$

$$H(R_i \cup R_j) = False \quad (25)$$



Figure III.7. Depicts a region segmentation result of an original image

3.4.1 Region-Growing Algorithm

This algorithm starts from some pixels representing distinct image regions and grows until they cover the entire image (Jun, 2010; Tremeau & Borel, 1997).

1) At each stage, k and for each region, $R_i(k)$, $i = 1, \dots, N$, Check for unclassified pixels in the neighborhood of each pixel of the region border.

2) Before assigning such a pixel x to a region $R_i(k)$, Check region homogeneity:

$H(R_i(k) \cup \{x\}) = TRUE$, is valid

3) The arithmetic mean $M1$ & $M2$ and standard deviation(sd) of a region R_i having n pixels of regions $R1, R2$ is calculated to take a merge decision.

if $|M1 - M2| < k * sd(R_i)$, $i = 1, 2$.

then two regions are merged

3.5 Edge-based segmentation

A set of linked pixels on the boundary between two regions, when there are intense discontinuities such as grey shift, colour distinctness, texture diversity, and so on, is referred to as an edge (Kang et al., 2009; Maini & Aggarwal, n.d.; Yogamangalam & Karthikeyan, 2013). Those discontinuities can be used to segment an image. With this technique, detected edges in an image are assumed to represent object boundaries and used to identify these objects. There are many ways to perform edge detection. If the level of detecting accuracy is

too high, noise may introduce artificial edges, outlining images unreliable; otherwise, if the degree of noise immunity is too high, some areas of the image outline may go unnoticed, and the position of objects may be incorrect.

3.7 Grid-based segmentation

This is a straightforward technique for segmenting an image. (S. L. Feng et al., 2004) A rectangular grid with fixed-size slides over (that can overlap) the image. An extracted feature is extracted for each rectangular grid. The rectangle dimension can be varied to create a multi-scale variant of grid partitioning (Lim & Jin, 2005). It is possible to cope with changes in object placements and image scale changes by combining overlapping and multi-scale partitioning. In annotation tasks, the rectangular grid outperforms the method based on region segmentation, according to (S. L. Feng et al., 2004). There is also a significant reduction in the amount of time it takes to segment the image.

Do not forget that there is a segmentation using deep learning; to get more information, browse the work (Minaee et al., 2021).

Table III.3: Comparison and Evaluation Of Segmentation Algorithms in different parameters

Parameter	Spatial information	Region continuity	speed	Computation complexity	automaticity	Noise resistance	Multiple object detection	Accuracy
Graph-based segmentation	ignored	good	moderate	expensive	automatic	moderate	fair	fine
Contour-based segmentation	considered	good	slow	less	interactive	moderate	fair	moderate
Clustering-based segmentation	considered	reasonable	fast	rapid	automatic	moderate	fair	moderate
Region based segmentation	considered	good	slow	rapid	Semi-auto	less	fair	fine
Edge-based segmentation	ignored	reasonable	moderate	moderate	interactive	less	poor	moderate
Grid based segmentation	considered	reasonable	fast	less	Semi-auto	moderate	fair	fine

4. Conclusion

In this chapter, we have briefly reviewed segmentation methods and extracted features used in AIA.

Even though low-level feature-based image annotation approaches aid in bridging the 'semantic gap' by providing abstract-anatomical descriptions for images, the results are frequently subjective and fall short of articulating semantic subtleties.

Semantic-oriented annotation approaches have been developed as an alternative. These methods aid in the generation and inference of semantic descriptions that reflect the semantic content of images, either in terms of visual qualities (regions, objects, etc.) or the spatial relationships between the items that show on the images. However, the semantic content of the image is only partially expressed.

Chapter IV: AIA Based on Machine Learning

1. Introduction	59
2. Unsupervised ML	60
2.1 Clustering	60
2.2 Hidden Markov Model (HMM)	60
2.3 Artificial Neural Network	62
3. Supervised-learning	64
3.1 The k-Nearest Neighbor (k-NN)	65
3.2 Decision Trees	65
3.3 Support Vector Machine	66
3.4 Naive Bayes	67
4. Deep learning	68
4.1 Deep Neural Network	68
4.2 Deep convolutional neural networks	69
5. Conclusion	74

1. Introduction

The most important data analysis method is machine learning (ML) (Adnan et al., 2019), which uses algorithms to learn from available data iteratively. Models programmed to accept new data are used to execute iterative follow-up. These models may be able to make significant predictions and decisions. In this chapter, we will explain some of the machine learning techniques that have been adapted into the automatic image annotation approach. ML aims to create computer systems that can learn and adapt to their environments (Carbonell, 1981; Dietterich, 2002; Dietterich & Oregon, 1996). The overall goal of machine learning is to improve the system's efficiency and effectiveness. Machine learning approaches can be classified into three categories: supervised ML, unsupervised ML, and deep learning. Figure IV.1 shows the diagrammatic representation of ML techniques.

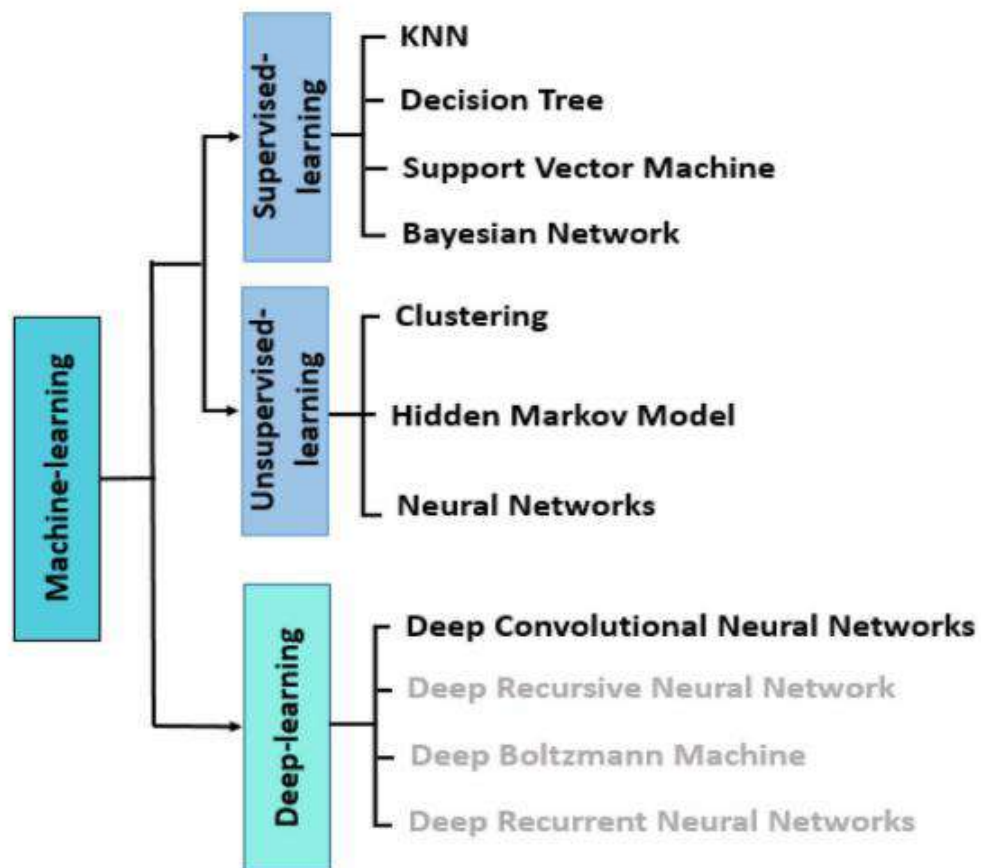


Figure IV.1 : diagrammatic representation of ML techniques (Bouchakwa et al., 2020)

2. Unsupervised ML

Unsupervised ML algorithms analyze input data, group data points based on perceived similarities, and draw inferences based on these similarities (Bouchakwa et al., 2020). Clustering, Hidden Markov Models (HMM), and Artificial Neural Networks (ANNs) are unsupervised learning approaches most often.

2.1 Clustering

Clustering (Agarwal, 2014; Arockiam et al., 2012; McGregor et al., 2004) is an unsupervised learning task that aims to find a finite number of clusters to characterize a set of data. The groups generated following the execution of a clustering process are called clusters. The technique of splitting an extensive data set of objects into small subgroups is the underlying concept of clustering. The clustering process has no training stage and is often used when the clusters are not known in advance. Indeed, the attributes providing the best clustering should be often identified in the first stage. Each small subset is a distinct cluster, with objects clustered together based on intra- and inter-class features. Clustering result quality may depend on similarity measures like Euclidean distance and Manhattan distance between two objects of numeric data used by the algorithm.

In theory, clustering is more efficient when intra-class similarity is maximized, and inter-class similarity is minimized. Things within clusters resemble those in the same cluster but are distinct from objects in other clusters. Similarity and dissimilarity are determined using object attribute values and distance metrics. The quality of a given clustering method is also computed according to its ability to discover some or all hidden patterns.

2.2 Hidden Markov Model (HMM)

The Hidden Markov Model (Eddy, 2004; Fine et al., 1998) is a finite state machine with a fixed number of states. It permits the provision of a probabilistic framework for modelling time series of multivariate observations. It consists of a statistical Markov model where the system being modeled is supposed to be a Markov process with unobserved (hidden) states. An HMM can be considered the most straightforward dynamic Bayesian network. In HMM, the state is not directly visible, but the output, which depends on the state, is visible. Each

state has a probability distribution on the possible output tokens. Then, the sequence of tokens generated by an HMM provides specific information about the sequence of states.

Ghoshal et al. (Ghoshal et al., 2005) have used an HMM for automatic annotation of images with keywords from a generic vocabulary of concepts or objects for content-based image retrieval by positing that an image is represented as having been generated by a hidden Markov model, whose states represent concepts, and that the image is represented as a sequence of feature- vectors describing low-level visual properties such as color, texture, or oriented-edges. The model's parameters are estimated from a set of manually annotated (training) images. Each image in an extensive test collection is then automatically annotated with the a posteriori probability of concepts present in it.

According to Wang et al. (J. Z. Wang & Li, 2002), humans tend to view images as a whole. As a result, some semantic notions can't be learned in a single region. The semantic indexing of images utilizing 2-D HMM for image annotation has also considered the relationships between regions.

Senthilkumar et al. (M Saleem, R Senthilkumar, 2015) introduced a method to annotate images with keywords from a generic vocabulary of concepts or objects for content-based image retrieval. The suggested method is based on HMMs for automatic and annotation-based image retrieval. A Semantic annotated Markovian Semantic Indexing (SMSI) is introduced in the automatic annotation task. It consists in modelling the images, represented as a sequence of feature vectors characterizing low-level visual features, like color, texture, and oriented edges, as having been stochastically provided by an HMM, whose states represent concepts. The model's parameters are estimated from manually annotated (training) images. Then, each image from an extensive test collection is automatically annotated with a posteriori probability of concepts present within it. Image Annotation Using Spatial HMM is a 2-D generalization of the traditional HMM in that both horizontal and vertical transitions between hidden states are considered. After annotating images, semantic retrieval of images can be performed by using a Natural Language processing tool, namely WordNet, and measuring the semantic similarity of annotated images in the database using Markovian Semantic Indexing (MSI)(Bhatt et al., 2010).

2.3 Artificial Neural Network

The Artificial Neural Network (ANN) (Bouchakwa et al., 2020) is an unsupervised machine-learning algorithm inspired by the human brain's natural way of information processing. It can learn from examples and provide decisions about new samples. ANN is credited for its ability to learn multiple classes all at once. An ANN consists of three layers: input, hidden, and output. Each layer consists of nodes (or neurons) performing numerical computations and other operations. Each neuron from a layer is interconnected with other neurons presented in consecutive layers. A bias is assigned to each layer, and a weight is assigned to each interconnection. Fig. 13 shows a simple neural network.

The input layer has neurons equal to the dimension of the input sample. It is responsible for receiving large volumes of data as inputs in different formats (text, images, CSV files, etc.). The output layer is responsible for producing the target outputs. All the calculations are performed in the hidden layer. Indeed, each neuron from the hidden layer operates as a processing element. It is governed by an activation function, which provides output according to the weights of the connecting edges and the outputs of the neurons of the previous layer. An NN learns the edge weights during the training process to minimize the overall learning error. To classify a new sample, each output neuron generates a confidence measure. The class corresponding to the maximum measure indicates the decision about the sample.

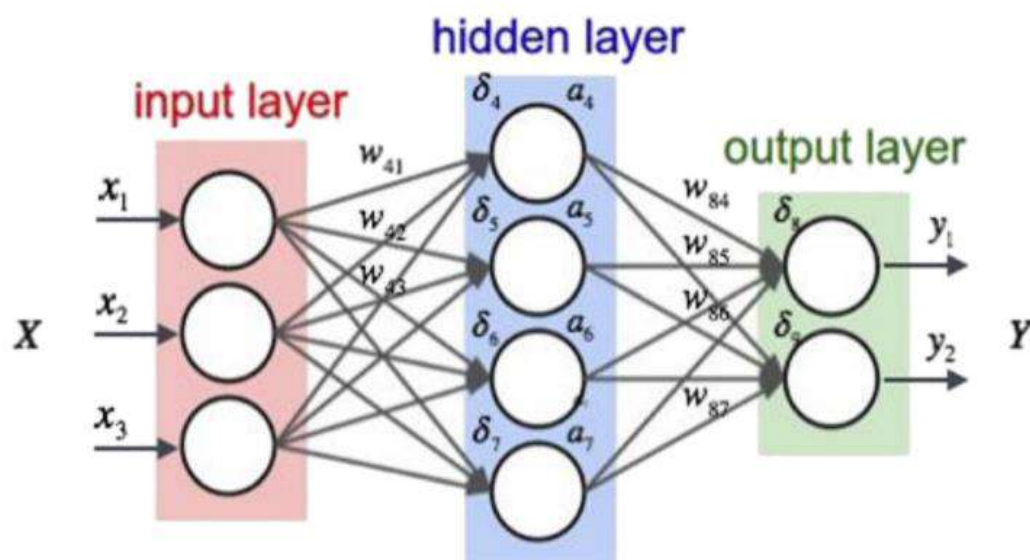


Figure IV.2. A simple Neural Network(Dutta, 2019)

Hambali et al. (Hambali et al., 2017) proposed a fruit classifier using a simple neural network model. The main aim of this study is to categorize 'jatropha fruits' according to their color features. The input layer consists of six neurons $\{x_1, x_2, \dots, x_6\}$, each representing the color of the elements; R, G, B, L^* , A^* , and B^* , respectively. The input layer receives the signal and then distributes the signal to the neurons in the hidden layer. The number of neurons in the hidden layer is seven, which is assumed sufficient to generate good prediction results. The output layer consists of four neurons, $\{t_1, t_2, t_3, t_4\}$, representing the quality that fruit can have; 'immature', 'under mature', 'mature', and 'over mature', respectively.

(Park et al., 2004) suggested a method of content-based image classification using a 3-layer ANN, where the hidden layer consists of 49 neurons. The images for classification are object images, which can be divided into background and foreground. Thus, a pre-processing step is proposed for segmenting an image into a set of regions. The largest region at the centre of the image is used to identify the image. The regions with similar color distribution to the central region are considered foreground (objects) regions. The foreground regions are used to extract the statistical texture features, which are transmitted to the ANN to classify the image into one of 30 concepts.

In their study, (Kuroda & Hagiwara, 2002) used four different 3-layer ANNs to classify image regions hierarchically. The numbers of neurons used in the hidden layers of these networks are 30, 10, 20, and 20, respectively. In this classifier, an image is first composed of some regions. Then each region is roughly classified into three broad categories, namely: 'sky', 'water', and 'earth', using SEW neural network. Second, the image features are extracted from each of the categories. The impression words (like 'bright/dark', 'heavy/light', 'warm/cool', 'emotional/reasonable' and 'rural/urban') are estimated from the image by using the second neural network called IW network. The regions belonging to the sky or earth categories are classified into much more complex objects, such as 'blue sky', 'cloud', 'sunset', 'mountain', 'green' and 'rock', using the OR neural network. The fourth neural network does not classify any region. Still, it permits associating an image with a vector of 18 dimensions, and each dimension measures the degree of certain global characteristics of the image, like 'bright/ dark', 'rural/urban', and 'busy/plain'.

3. Supervised-learning

Supervised learning is the research of algorithms, which reason from externally supplied instances to produce general hypotheses that constitute predictions about future instances. In other words, the SL aims to build a concise model of the class label distribution in predictor features. The resulting classifier is after that used to assign class labels to the testing instances, where the predictor feature values are known, but the class label value is unknown. Figure IV.3 illustrates the process of SL gradually.

SL is the most used technique in applications where available past data predict the expected future events. Equation (26) shows the general representation of SL as:

$$D = \{(x_i, y_i)_{i=1}^N, \quad X_i(x_i^1, x_i^2, \dots, x_i^d)\} \quad (26)$$

Where D is the training dataset, N is the number of training examples, X_i is the attributes set, and y_i is the categories assigned to X_i .

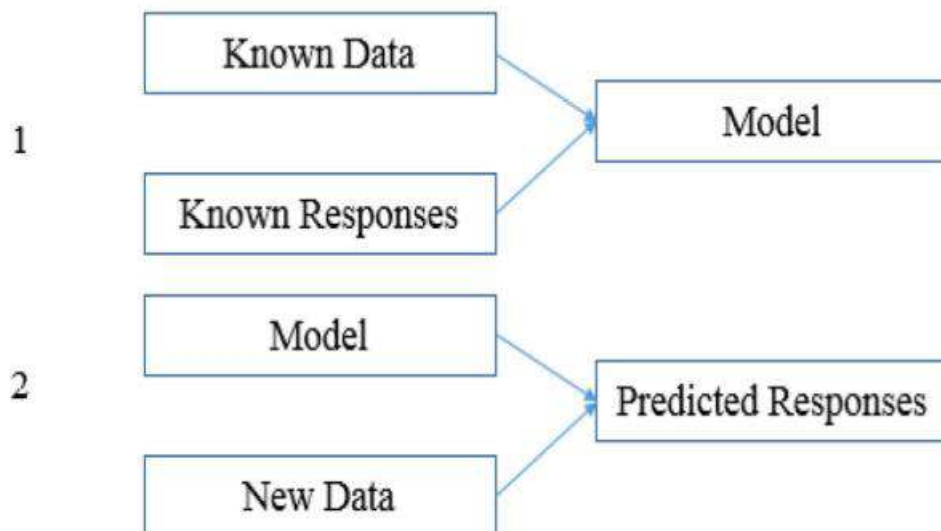


Figure IV.3: flowchart of the supervised learning process (Bouchakwa et al., 2020)

3.1 The k-Nearest Neighbor (k-NN)

KNN is a non-parametric supervised method based on similarity in the feature space. In its simplest form, the label of an unknown test sample is assigned by the majority vote of its K-nearest neighbours from the training data (whose labels are known). If $K = 1$, the test data is simply assigned the class of the single nearest neighbour. The assignment for $K=1$ and $K=4$ is shown in Figure IV.4

The performance of KNN depends on the value of the hyperparameter K and the distance metric used. If the value of K is minimal, the test sample ends up with a small neighborhood, and this could result in poor performance because of sparse, noisy, ambiguous or poorly labeled data. If we try to increase the value of K, it introduces outliers from other classes. Advanced KNN algorithms also use various weighting schemes to assign weights to the neighbours' contributions, so the nearer neighbours contribute more to the majority vote than the distant ones. For example, a standard weighting scheme can give each neighbour a weight of $1/d$, where d is the distance to the test sample.

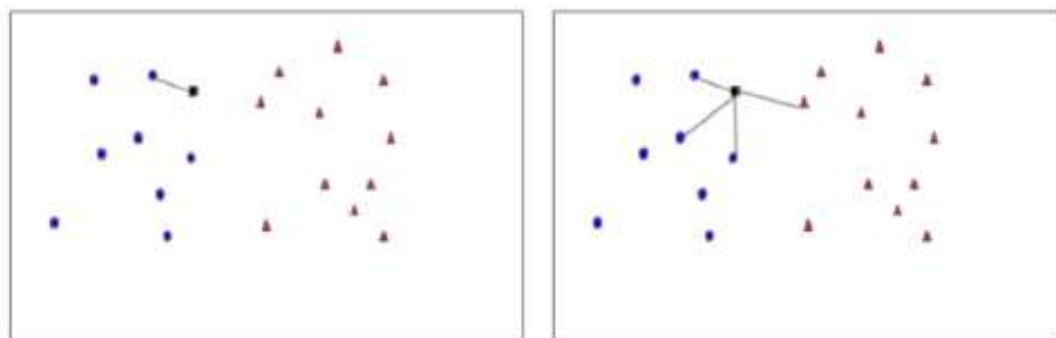


Figure IV.4: An example of K-nearest neighbour assignment with $K = 1$ (left) and $K = 4$ (right) (best viewed in colour).

3.2 Decision Trees

The decision tree (Quinlan, 1986) is a natural way of presenting a decision-making process because of simple and easy for anyone to understand. A decision tree is a simple but powerful form of multiple variable analysis used for attribute values to class mappings. Decision trees can be used in place of multiple linear regressions such as statistical form analysis or in intelligent business systems where multidimensional data analysis is expected (Lomax & Vadera, 2013). A tree is a leaf node labeled with a class or a structure consisting of a test node

linked to two or more subtrees. A test node computes some outcomes based on the attribute values of an instance, where each possible outcome is associated with one of the subtrees.

An instance is classified by starting at the root node of the tree. If this node is a test, the outcome for the instance is determined and the process continues using the appropriate subtree. When a leaf is eventually encountered, its label gives the predicted class of the instance (Quinlan, 1996). Figure IV.5 shows the classification process using decision tree (DT) classifier

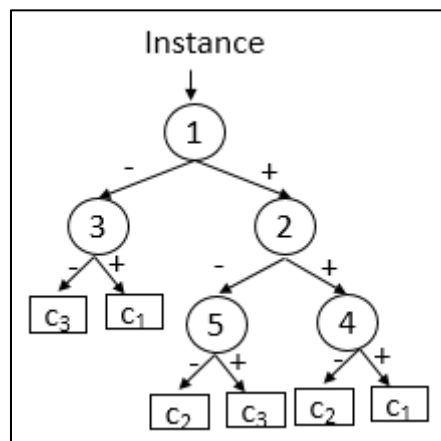


Figure IV.5. Design of Decision Tree (DT) classifier.

3.3 Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier defined by a separating hyperplane with a maximum margin. In other words, given labeled training data, the algorithm outputs a hyperplane that optimally separates the data according to their labels (positive/negative) with maximum margin. Let us assume a training set $\{x_i, y_i\}_{i=1}^n$ of n examples, where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$ denotes whether it belongs to the class or not. In SVM, the goal is to learn a hyperplane, characterized by the parameters w and b , such that the following constraints are satisfied for all data points:

$$w \cdot x_i + b \geq 1, \text{ if } y_i = 1 \quad (27)$$

$$w \cdot x_i + b \leq -1, \text{ if } y_i = -1 \quad (28)$$

These constraints can be re-written as:

$$y_i(w \cdot x_i + b) \geq 1 \quad (29)$$

Since this is a hard constraint, we can approximate it by introducing non-negative slack variables ξ_i .

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \forall i \quad (30)$$

This leads to the following optimization problem:

$$\min \frac{\gamma}{2} \|w\|^2 + \frac{1}{n} \sum_{i=1}^n \xi_i \quad (31)$$

$$s. t. y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \forall i \quad (32)$$

Here, $\|\cdot\|^2$ is the squared L_2 norm that acts as a regularizer on w and ensures a hyperplane with

the maximum margin separation between the two classes. $\lambda > 0$ is a hyperparameter that handles the trade-off between the regularization term and the loss function (also called hinge-loss) that penalizes the violation of the constraints. Given a new sample x , we predict whether it belongs to the given class or not based on the $y = \text{sign}(w \cdot x_i + b)$. SVM provides several practical advantages, such as a convex optimization problem, extensive margin guarantees, good generalization, scalability, and fast testing time, and is often used as the de facto baseline in classification tasks

3.4 Naive Bayes

(Ivasic-Kos et al., 2016) suggested a two-phased multi-level picture annotation methodology. The first phase involves using an NBC to classify low-level image features. In contrast, the second phase involves using a fuzzy Petri net-based knowledge representation scheme to expand the vocabulary level and incorporate multi-level semantic concepts related to images into image annotations. Incorporating clustering with pair-wise constraints for AIA, Rui et al. (2005) and Heller and Ghahramani (2006) presented a semi-NBC method. Experiments have demonstrated that the strategy improves annotation performance significantly.

Darwish et al. (2016) present a novel approach to picture annotation that is multi-instance and multi-label (MIML). Images are initially segmented using the Otsu method, which

maximizes the image's intracluster variance to determine an ideal threshold. The Otsu method is tweaked with FA to reduce runtime and improve segmentation accuracy.

(2014) proposed a Bayesian framework for image retrieval based on content. The advantage of this method is that it uses numerous photos to conduct retrieval rather than just one. Based on the Bayesian criterion and the marginal likelihood of discovering the photographs most likely belong to a query group, the method produced good results. Based on the results thus far, it seems clear that the semi-Naive Bayes is more effective than the NB. When constructing regions acquired by latent topic allocation, remember that a Bayesian learning model is used. Meanwhile, compared to other machine learning models, this is highly sophisticated. The system's reliability may be difficult to determine due to numerous conditional probabilities.

4. Deep learning

4.1 Deep Neural Network

A DNN is a neural network with more than two layers characterized by a specific level of complexity. For complex data processing, the DNNs relies on extensive mathematical modelling.

Zhu et al. (2015) proposed a new multimodal deep learning network framework for learning intermediate representations and ensuring proper network initialization. The distance metric functions on each modality were then optimized via backpropagation. Ultimately, the exponentiated gradient online learning technique was used to maximize the combinational weights of different modalities. Additional deep learning research is required to determine the number of feature dimensions required to achieve satisfactory system performance for a given neural network framework. Another factor to examine is the mechanism employed to improve the resilience of specific deep learning architecture.

Yang et al. (2015) developed a novel MVSAE Model for automatically establishing the correlations between high-level semantic keywords and low-level image characteristics. The SAE was first altered using an iteration technique and a sigmoid function predictor. The image keywords were then solved using an unequal distribution. At varying levels of keyword frequency, the imbalance learning method has distinct effects. Because a low-frequency

keyword tends to cause a high-frequency keyword to be misclassified, the F1 score drops slightly towards high-frequency keywords. The low-level frequency keywords, on the other hand, perform better than the original SAE. Yang et al. (2015) suggested a Multi-View Stacked Auto-Encoder (MVSAE) framework for determining the connections between high-level semantic information and low-level visual information.

4.2 Deep convolutional neural networks

Convolutional Neural Network (CNN) is a feed-forward artificial neural network with learnable weights and biases that can operate on input volumes such as multi-channelled images. Inspired by biological processes, its connectivity pattern between neurons is analogous to the organization of the animal visual cortex. While artificial neural networks have been in use for various tasks, it was CNN that first demonstrated the capability of incorporating a large number of hidden layers in a network. Since then, it has led to massive growth in the ability to solve Computer Vision problems. A CNN is typically a sequence of layers with neurons that transform one volume of activations to another through a differentiable function starting from the raw image pixels on one end to class scores at the other. The neurons in a layer are connected to a small region before it, except at the fully connected layer(s) towards the end of the network. The parameters of the network are end-to-end trainable using a loss function on the output of the last layer. A CNN architecture (Figure IV.7) mainly consists of an input layer, a stack of convolutional, ReLU and pooling layers, and finally, the fully-connected layers, as described below.

Input Layer: Input is an image of dimension height \times width \times depth, containing raw pixel values and depth denoting the three color channels (R, G and B).

Convolutional Layer: The convolutional layer is the core building block of a CNN, consisting of a set of learnable parameters called filters, which are 3-dimensional volumes small spatially (along width and height) but with the same depth as the input volume (Figure IV.7). Every filter is convolved spatially across the input during the forward pass to compute dot products between the filter (weights and biases) and the local input region at any spatial position of the input volume. This produces a 2-dimensional activation map showing filter responses at different spatial positions. In other words, the network learns filters that activate when it detects some specific type of feature (e.g., edges of some orientation, specific

patterns, etc.) at some spatial position in the input. The output volume of the convolution layer is the activation map for all filters stacked along the depth dimension.

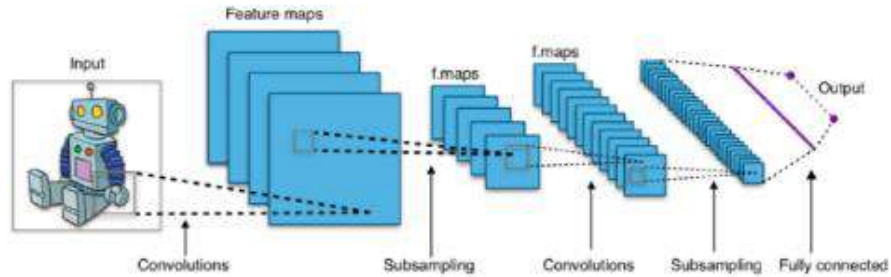


Figure IV.6. Example of typical convolutional neural network architecture(Moutarde, 2019).

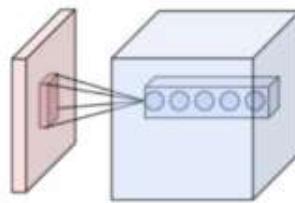


Figure IV.7. The connection between neurons of the convolutional layer (blue) and the input volume(red)(Dutta, 2019)

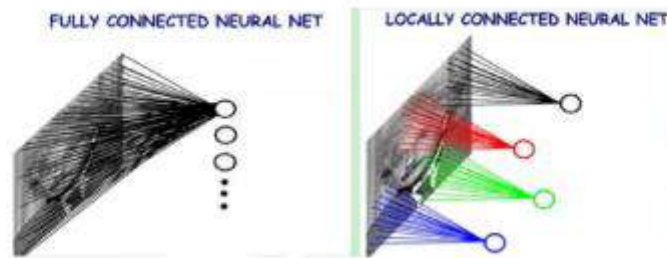


Figure IV.8. Local connectivity of convolutional layer



Figure IV.9. Examples of pooling layers(Dutta, 2019)

Pooling Layer: The pooling layer performs a downsampling operation on the input along its spatial dimensions (width, height) to progressively reduce the spatial size of the representation in the network. It is done independently on every depth slice of the input, thus keeping the depth dimension unchanged. Pooling reduces the number of network parameters, controls overfitting, and provides a form of translation invariance. The most common forms of it used are max pooling and average pooling. Max pooling replaces the input region it is connected to with the maximum value, whereas average pooling replaces it with the input region's mean (or average) value (Figure IV.9).

ReLU Layer: The ReLU layer applies an element-wise activation function $\max(0, x)$ that thresholds the neurons' outputs at zero. This layer adds non-linearity to the CNN.

Fully-Connected Layer: As the name implies, each neuron in a fully-connected layer is connected to all of the neurons in the previous layer (Figure IV.8). Generally, for the classification task, the final fully-connected layer of any CNN consists of C hidden units, where C is the number of classes. The output of the C classes is passed through a softmax or sigmoid activation function to obtain class-probability scores corresponding to each class.

Table IV 1 : A comparison between methods and present some of the Advantages and Disadvantages of each model

Algorithm		Advantages	Disadvantages
Unsupervised-learning	Clustering	<ol style="list-style-type: none"> 1. Simple and relatively scalable, 2. Appropriate for datasets with compact spherical clusters, which are well-separated. 3. Embedded flexibility concerning the level of granularity, 4. Well adapted for problems that involve point linkages, such as taxonomy trees. 	<ol style="list-style-type: none"> 1. Serious effectiveness degradation in high dimensional spaces. 2. Poor description for clusters. 3. Requires a manual specification of the number of clusters in advance. 4. High sensitivity to initialization phase, outliers and noise. 5. Frequent entrapments in the local optima. 6. Inability to perform corrections once the splitting or merging decision is made, 7. Cloudiness of termination criterion, 8. Expensive for massive and high dimensional datasets, 9. Serious effectiveness degradation in case of high dimensional spaces
	HMM	<ol style="list-style-type: none"> 1. Allows an efficient learning that can be performed directly from raw sequence data. 	<ol style="list-style-type: none"> 1. Not completely automatic and requires training using annotated data, 2. The size of training data can be an issue.
	ANN	<ol style="list-style-type: none"> 1. Enables to manipulate non-parametric training data, 2. Capability to present functions, such as AND, OR and NOT, 3. Consists of data driven self adaptive technique, 4. Efficiently handles noisy inputs, 5. Computation rate is important. 	<ol style="list-style-type: none"> 1. Semantic poverty, 2. Problem of over-fitting, 3. The training of ANN is time consuming, 4. Difficult to define the network architecture.
supervised-learning	KNN	<ol style="list-style-type: none"> 1. Manipulate non-parametric training data, 2. Training step is very fast, 3. Simple to learn, 4. Robust to noisy training data, 5. Effective when training data is large. 	<ol style="list-style-type: none"> 1. Biased by the value of k, 2. Computation are complex, 3. Limitation of the memory, 4. Testing step runs slowly.

	DT	<ol style="list-style-type: none"> 1. Manipulate non-parametric training data, 2. Does not required an extensive training 3. Generates the deep learning features hierarchical associations between input variables to predict class membership and produces a set of rules that are easy to interpret, 4. Simple and efficient computational 	<ol style="list-style-type: none"> 1. The computation becomes complex when various outcomes are correlated and/or vari- ous values are undecided.
	SVM	<ol style="list-style-type: none"> 1. Achieves optimal class boundaries by finding the maximum distance between classes, 2. Provides a good generalization capability, 3. The adjustment problem is eliminated, 4. Computational complexity is reduced, 5. Simple to manage the error frequency and decision rule complexity. 	<ol style="list-style-type: none"> 1. Result transparency is weak, 2. Training step is time consuming, 3. Structure of the algorithm is difficult to understand, 4. Determination of optimal parameters is complex when there is non-linearly separable training data.
	NB	<ol style="list-style-type: none"> 1. Performance is good, 2. Easy to implement, 3. Takes less computational time for processing. 	<ol style="list-style-type: none"> 1. The dependencies existing between variables are ignored, which would cause it to provide less accurate predictions.
Deep learning		<ol style="list-style-type: none"> 1. Treats large data, 2. Process complicated relationships, 3. Derives robust characteristics, 4. No manual choice is needed, 5. Multi-labeling of images. 	<ol style="list-style-type: none"> 1. Optimum is local, 2. Training stage cannot be controlled, 3. Needs large training images.

5. Conclusion

In the chapter, we have focused on identifying the parameters of the image annotation systems based on machine/deep learning. Deep learning (DL) is a particular type of Machine-learning (ML), which is also a subfield of Artificial-intelligence (AI). After that, many Machine-Learning algorithms were introduced to automate the annotation process and reduce human efforts. We presented the AIA methods based ML and Support Vector Machine (SVM) based AIA, k-Nearest Neighbor (kNN) based AIA, Deep Neural Network (DNN) based AIA and Bayesian-based AIA. A comparison of the A types of AIA approaches has been presented based on the underlying idea, the feature extraction method, annotation accuracy, computational complexity, and datasets. However, the main issue in ML is that an inadequate data representation often degrades the quality of the results and leads to lower performance than a suitable data representation.

Chapter V: A framework for AIA

1. Introduction	76
2. Approach	77
2.1 Image segmentation using algorithm JSEG	77
2.2 Region representation	78
2.3 Feature aggregation	79
2.3 Calculating Blob/Label co-occurrences:	80
2.4 Annotating new images	81
3. Conclusion	82

1. Introduction

The sections in this chapter present the used methods for each component in the system Figure V.1 with the motivations behind each choice of method. The main objective is to assign a set of labels for a given image, each representing one region (object) within the image. KNN regression has been employed to enhance the representation of regions in the input feature space and the propagation of labels in the output semantic space.

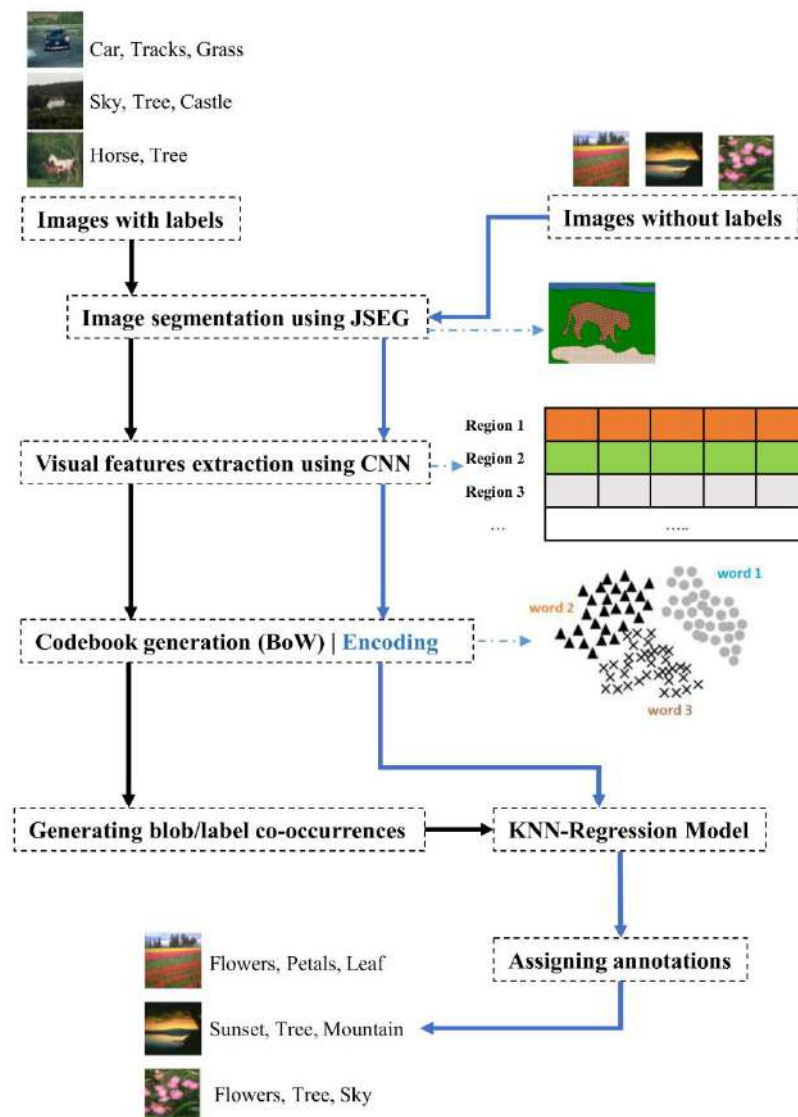


Figure V.1: Detailed architecture model of the system showing all components and the data flow in between, from the data sources to the annotated images. Black solid arrows correspond to training images whereas the blue ones correspond to test images.

2. Approach

Conventional AIA algorithms consider the image as holistic by analyzing images global rather than dealing with each present object. In real cases, however, few concepts may describe the image holistically such as ‘joy’ or ‘wild’, but most concepts are concerned with some specific regions (areas) of the image such as ‘football’, ‘human’, ‘cloud’, etc. As a result, for an AIA system to produce good annotation results, it must account for visual distinctions across regions as well as semantic interconnections between labels. Given that a concept-region co-occurrence matrix is derived from an annotated training image subset, our proposed solution investigates the similarity among characteristics of a candidate region and the training subset using this concept-region co-occurrence matrix. By doing so, we ensure that visual correlations among areas are taken into consideration. Thereafter, we employ a k-nearest neighbors regression (knn-r) algorithm to annotate new regions. Figure V.1 depicts a general scheme of the proposed approach.

As the scheme in Figure V.1 shows, our model takes a set D of images $D=\{I_1, \dots, I_N\}$ some of which are labeled (for training) and the rest of which are not. It should be mentioned that each training image I_n is labeled with I_{cn} concepts: $I_{cn} \in C / C = \{C_1, \dots, C_M\}$. All images are passed through a preprocessing step in which they are segmented, using the JSEG algorithm, into visually homogeneous regions. An aggregation approach has been subsequently used to decrease a large number of areas by codifying comparable areas into blobs (codebook) with each blob corresponding to one label. Using the generated codebook and the annotation from the training subset, our model generates a co-occurrence matrix that codifies the appearance frequency of each blob/concept. Finally, we engage knn-regression to predict annotations corresponding to blobs extracted from unannotated images. Each of these steps will be further discussed hereafter.

2.1 Image segmentation using algorithm JSEG

According to (Bhagat & Choudhary, 2018), the best way to recognize objects from an image is to segment them and then extract features from those segmented regions. However, object segmentation, both using supervised and unsupervised approaches, is itself a complex task. Despite the difficulty of achieving precise and accurate semantic segmentation, it has

been proven, on many occasions, those segmented areas are valuable and effective annotation cues (Darwish, 2016; J. Zhang et al., 2016).

JSEG is a powerful unsupervised segmentation algorithm for color images that proved its effectiveness and robustness in a variety of applications (Khattab et al., 2014; Yining Deng et al., 1999). The widely used JSEG algorithm is a region growing approach. JSEG has recently witnessed various improvements to improve its performances such as the problem of over-segmentation (Aloun et al., 2019; Yining Deng et al., 1999). In our study, the JSEG proposed in (J. Zhang et al., 2018) has been employed to segment the image into a set of semantic regions as illustrated in Figure V.2

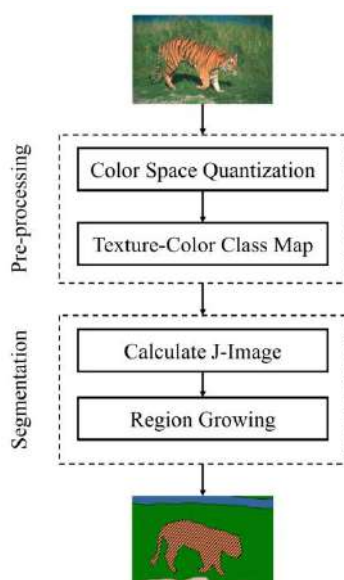


Figure V.2: a General scheme of Texture-enhanced JSEG (T-JSEG) segmentation method

2.2 Region representation

In region-based techniques, the visual characteristics of the image, such as color, texture, and form, are typically extracted from each region. Using local features instead of global ones has been proven to be more effective in image annotation tasks. Nevertheless, appropriate features must be selected to represent the essential substance of the image. For the task of image representation, deep CNNs have recently been shown to outperform, by a significant margin, state-of-the-art solutions that use traditional hand-crafted features. In our study, the learning transfer of off-the-shelf features extracted from a pre-trained CNN model has been used to represent the content of each image region. Learning transfer has shown high

efficiency in extracting visual features and demonstrated that features with sufficient representative strength can be extracted from the last layers (Oquab et al., 2014; Zeiler & Fergus, 2014). We have opted for a pretrained model for two reasons, the first one is we don't have a sufficient amount of data nor the necessary resources to train a new CNN model, the second reason is to speed up the training process of our model. MobileNet (Howard et al., 2017) model, shown in Figure V.3. has been adopted in the present work since it has proved high performance (both accuracy and rapidness) in many learning transfer-based methods.

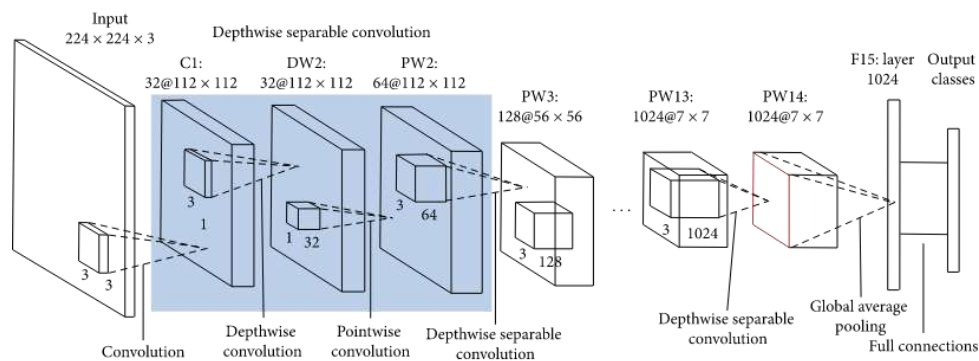


Figure V.3: MobileNet Architecture(W. Wang et al., 2020).

2.3 Feature aggregation

The JSEG algorithm doesn't necessarily generate an equal number of regions per image. Thus, extracting features from each region usually results in image descriptors with different sizes. To normalize the sizes of image descriptors, an aggregation method is generally utilized to produce a codebook that is used later on to codify the descriptors into equal size descriptors (Lai et al., 2020).

Vector of Locally Aggregated Descriptors (VLAD) is one of the most powerful aggregation techniques that's used to produce fixed-length vectors from local feature sets $X_i = \{x_j \in \mathbb{R}^F, j = 1, \dots, N_i\}$ having different sizes, where N_i is the number of local descriptors extracted from image i . VLAD generates, from the training set, a codebook $C = \{c_i \in \mathbb{R}^K, i = 1, \dots, M\}$ where M is the number of estimated clusters and c_i are their respective centers. Thereafter, a sub-vector v_i is obtained via accumulating the residual errors over an image X_i for each $i = 1, \dots, M$ (Lai et al., 2020).

$$v_i = \sum_{x_j: g(x_j, C) = c_i} x_j - c_i \quad (33)$$

where $g(x_j, C) = \operatorname{argmin}_{c_i \in C} \|x_j - c_i\|^2$ maps a descriptor x_j to its nearest cluster c_i . The descriptor D_i of the image X_i is a matrix of size $M \times F$ which is produced by concatenating all the corresponding codes $D_i = [v_1^T, v_2^T, \dots, v_M^T]$. This descriptor is power-normalized, and then l_2 -normalized, i.e.,

$$v_l := |v_l|^{0.5} \cdot \operatorname{sign}(v_l) / \|v\|_2, \quad l = 1, \dots, M. \quad (34)$$

The overall encoding process can be summarized as a function F that maps a codebook and a feature set to a global vector $v = F(X, C)$.

2.3 Calculating Blob/Label co-occurrences:

After having images segmented and descriptors extracted from regions, a clustering process must be performed to define local manifolds constituting the feature space. To this end, we employ the recent deep-clustering N2D algorithm. N2D learns an autoencoder embedding model and then searches this further for the underlying manifolds. Thereafter, a shallow network, rather than a deeper one, is used to perform clustering. N2D suggests that local manifolds learned on an autoencoder embedding are effective for discovering higher quality clusters. (McConville et al., 2020).

In our new space, image regions that are visually similar lie within the same manifold. Let's suppose that N2D has produced a set of clusters $C = \{c_1, c_2, \dots, c_M\}$ and the respective set S of label subsets s_i : $S = \{s_1, s_2, \dots, s_M\}$, then, an image that contributes by at least one region into the cluster c_j must contribute all of its labels to s_j . In other words, s_j holds labels from images that have at least one region in the cluster c_j . By exploiting both S and R , we can extract some useful complex semantic cues that link region-region, region-concept, and concept. To do so, we extract a concept-cluster co-occurrence matrix M in which each cell $M(c_j, r_i)$ indicates the appearance frequency of a concept (iow. label) l in the cluster, given the label subset s_i .

$$M(c_j, r_i) = \frac{\sum_{s_{ij} \in s_i} \delta_{s_{ij}, c_i}}{\|S\|} \quad (35)$$

where δ is the Kronecker delta function, and $\|S\|$ is a normalizer which represents the total number of labels that correspond to all the clusters.

The co-occurrence matrix M can be considered as a relatedness metric that measures the correlation among concepts and clusters. M will, later on, be used to calculate the conditional probabilities.

2.4 Annotating new images

Let's suppose that we have a new input image I_{new} without labels, and we want to assign annotations to it. Similarly, T-JSEG algorithm will be employed to segment the image I_{new} and produce a set of regions $\tilde{r} = \{\tilde{r}_1, \dots, \tilde{r}_s\}$. Since we have assumed that each region \tilde{r}_i corresponds to one annotation c_i from the annotation space, then we must calculate the conditional probabilities $P(c_i/\tilde{r}_i)$ to find out the best annotation that fits the region.

To assign a set of annotations, we perform a knn-regression while maximizing a Bayesian probability as follows:

1. Embed \tilde{r}_i descriptor into the appropriate manifold using the trained autoencoder model from N2D.
2. Retrieve k-nearest clusters using a simple Euclidean distance $C_{ri} = \{c_1, c_2, \dots, c_k\}$ and calculate, for each annotation a_i in the dataset, a regression probability: $P(a_i) = \sum_{l=1}^k M(a_i, c_l)$. This regressed value will be considered as a representative of the region \tilde{r}_i .
3. Maximize the following bayesian probability: $\arg_{a_i \in A} \max P(\tilde{r}_i) = \frac{P(\tilde{r}_i | a_i) P(a_i)}{P(\tilde{r}_i)}$, where $P(a_i) = \frac{1}{\text{Number of annotations}}$, and $P(\tilde{r}_i) = \frac{1}{k} \sum_{l=1}^k g(c_l)$, $g(c_l)$ calculates the center of the cluster c_l .
4. Assign the top fit concepts $C^* = \{a_j\}$ to the input image.

The rationale behind involving a neighborhood of clusters, rather than one cluster, to annotate one region is to ensure that we are taking into account information about blob-to-blob relationships, which grants higher error tolerance.

3. Conclusion

We have presented through this chapter the different contributions made in the thesis. Essentially, they are categorized into two significant contributions. The first contribution of the thesis concerns the study of the application of the image annotation models. The study of the second subject mainly includes the application of the We propose a novel image annotation approach, that is, Automatic image annotation using KNN regression. We aim to find the correct label (unique) for each region. Other techniques have been involved: segmentation using JSEG, feature extraction using CNN, and feature aggregation using the N2D algorithm.

Chapter VI:

Experiments and result analysis

1. Introduction	84
2. Experiment Setup	84
2.1 Datasets	84
2.2 Evaluation Metrics	85
2.3 Scenario 1: parameters tuning	86
2.4 Scenario 2: Comparing our method to the state of the art.	89
.5 Scenario 3: Computing cost	97
3. Conclusion	98

1. Introduction

This chapter is devoted to proving the efficiency of the proposed scheme across three scenarios. In the first scenario, we examine the impact of altering the parameters' values of our algorithm and try to tune them. In the second scenario, a comparison against state-of-the-art is conducted trying to demonstrate the superiority of our proposed algorithm. Finally, we investigate the complexity of our proposal by estimating the time consumed in the annotation process.

2. Experiment Setup

All experiments in this section have been carried out using the following configurations:

2.1 Datasets

We have used two well-known datasets, namely, Corel 5K and MSRC v2.

Corel 5K: It is a publicly available dataset that's commonly used for the task of image annotation. It is composed of 5000 images from 50 photo stock CDs annotated with 374 labels in total. Each CD includes 100 images on the same topic, annotated with 1 to 5 keywords per image. Due to the unbalanced nature of label distribution over images, most previous works consider using a few numbers of concepts (i.e., a subset of images) that appear frequently. However, we evaluate our proposed algorithm on both subset and complete datasets to prove its effectiveness and tolerance to the problem of unbalanced label distribution. Corel 5K is already splitted into train and test subsets comprising 4500 and 500 images respectively.

MSRC v2: This dataset contains 591 images grouped into categories having 23 concepts, each image explained using 1-7 keywords. MSRC v2 is splitted into train and test subsets comprising 394 and 197 images respectively.

Table VI.1 lists the essential characteristics of the used two datasets.

Table VI.1 : specifications of the used two datasets, Corel 5k and MSRC v2.

	Corel 5k	MSRC v2
Dataset size	5000	591
Train set size	4500	394
Test set size	500	197
Number of labels	371	23
Mean labels per image	3.4	2.5
Mean images per label	58.6	28.15

2.2 Evaluation Metrics:

To evaluate the performance of the proposed scheme, four widely known metrics for image annotation tasks have been opted for, namely: precision (P), recall (R), F1-score (F1) and N+. The formulas to calculate these three quantities are given respectively by the following equation.

$$P = \frac{1}{|S|} \sum_{s \in S} \frac{|\text{images annotated correctly with label } s|}{|\text{images annotated with label } s|} \quad (36)$$

$$R = \frac{1}{|S|} \sum_{s \in S} \frac{|\text{images annotated correctly with label } s|}{|\text{images having label } s \text{ in the ground truth}|} \quad (37)$$

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (38)$$

$$N+ = \text{the number of concept assigned correctly at least once} \quad (39)$$

It must be mentioned that region features are extracted from the final fully connected layer of the CNN model. This is due to the fact that information collected from the final FC layer is more suited to characterizing areas, especially when there is no stable color distribution (i.e., objects rather than textures)

2.3 Scenario 1: parameters tuning

The aim of this first scenario is to tune the values of our method's parameters which ensure sufficient performance. We firstly tune the most suitable aggregation method among the three well known methods: Bag of Visual Words(BoVW), Vector of Linearly Aggregated Descriptors (VLAD), and Fisher Vector (FV). Figure VI.1 represents the precision/recall yielded using features encoded by each of the aforementioned aggregation methods.

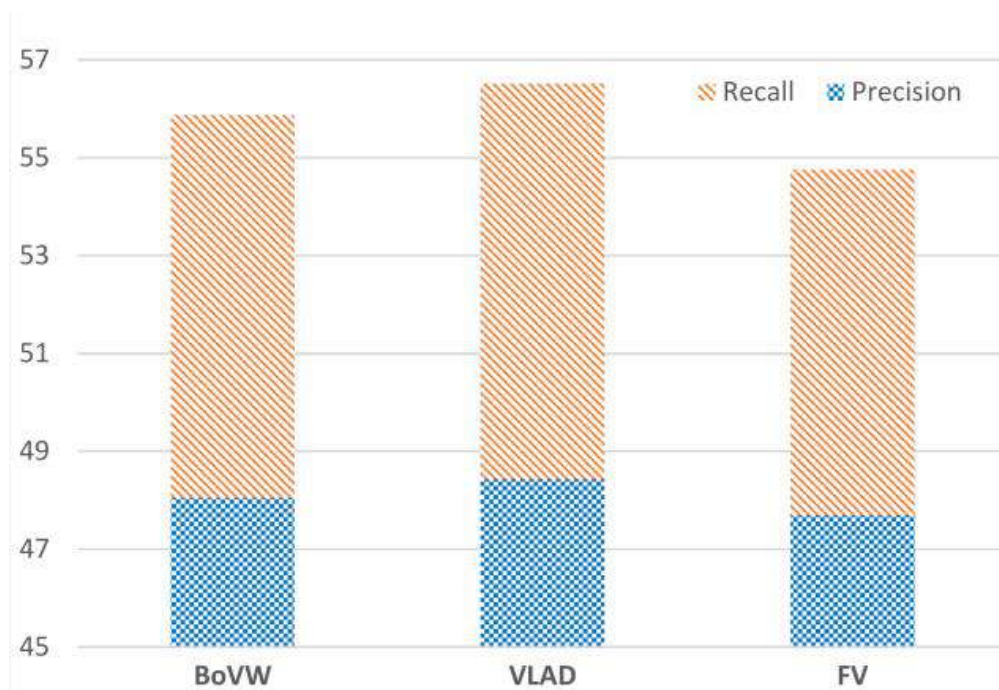


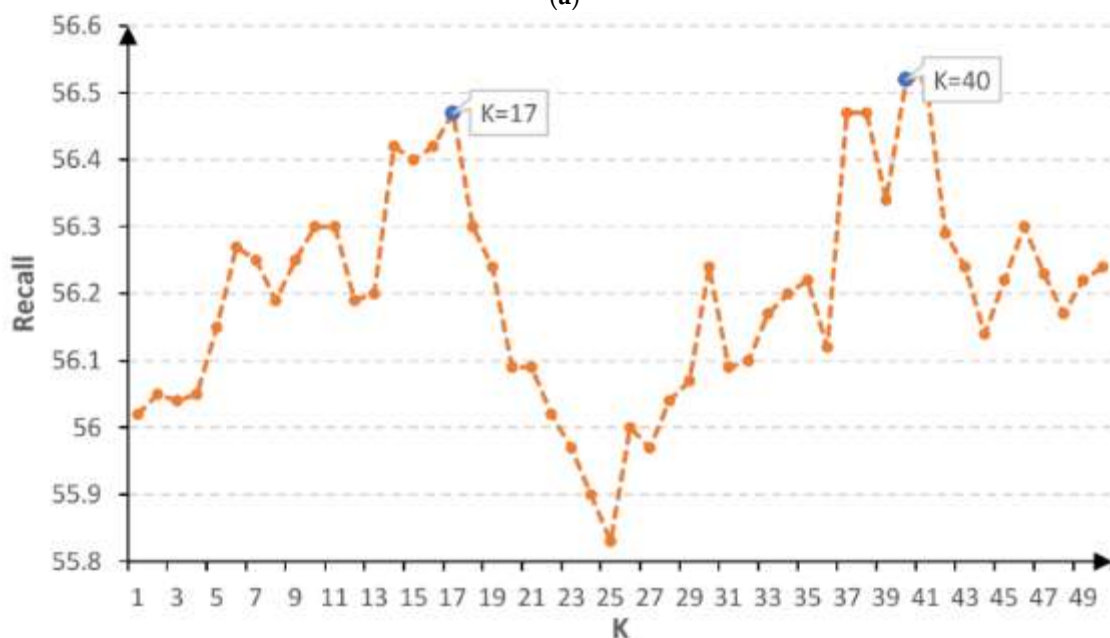
Figure VI.1: precisions/recalls yielded using the three aggregation methods : BoVW, VLAD and FV.

From Figure VI.1, it appears that VLAD has the best performance among the others. FV, on the other hand, has yielded the worst performance due to the second-order information it takes into account which is not helpful in cases of segmented homogeneous regions. We opted for VLAD in the remainder of this section because of the sufficient performance and the fast vector quantization it provides.

The K parameter of the KNN regression algorithm might be affected by different factors such as, the task it is used for, the length of the feature vector, the number of classes, etc. In order to determine which value fits most our task of automatic image annotation, we have evaluated the KNN algorithm with k values ranging from 1 to 50. Figure 0.2 shows the impact of changing k values' on the final precision and recall.



(a)



(b)

Figure VI.2: The impact of changing the value of k of KNN regressor on (a) the precision and (b) recall of our proposed method.

From Figure VI.2, it appears that our method grants the best performance at $K=40$. However, $k=17$ has rather been chosen to provide a trade off between precision and computation speed.

Since our method engages off-the-shelf CNN-based features, we evaluate several CNN models to determine which is the best for our task. The performance is not determined only in terms of precision and recall, but also in terms of time consumed in images processing. Figure VI.3 shows the impact of using different CNN models on the precision/recall of our proposed method.

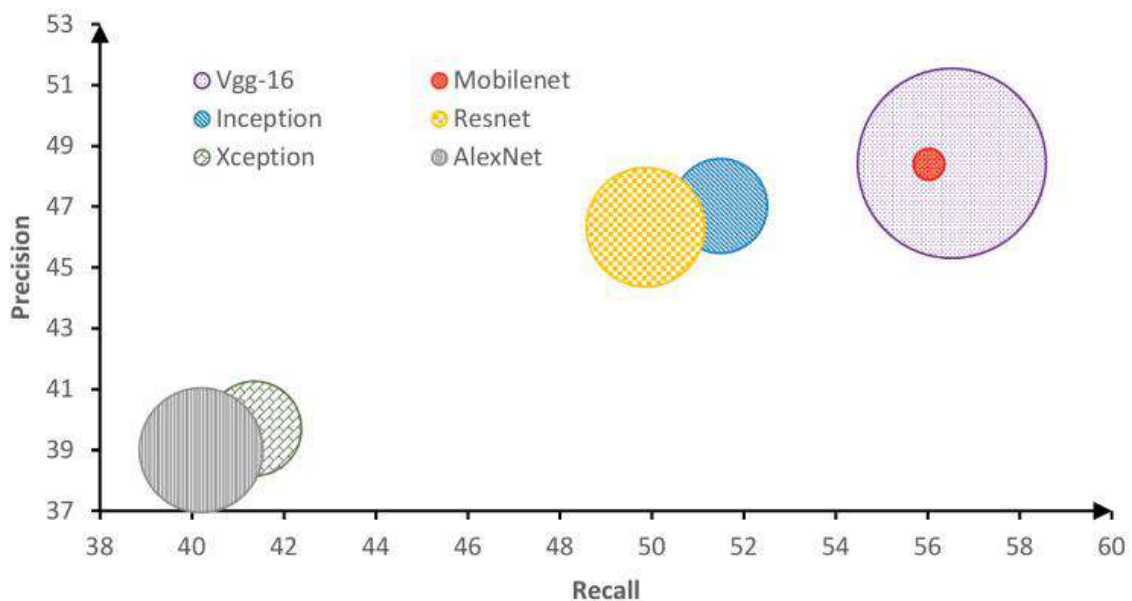


Figure VI.3: The impact of using different CNN models on our proposed method. The impact is measured in terms of precision, recall and complexity.

From Figure 0VI.3, it appears that the best two CNN models are Vgg-16 and MobileNet respectively. However, the latter suffers from the high complexity (huge number of parameters) which requires far more time of calculation (30 times slower) compared to the former. In our model, we have opted for MobileNet to achieve a better trade-off between accuracy and computation time.

In this first scenario, we aimed at tuning parameters to get, to some extent, satisfactory results. Thus, VLAD aggregation method, $k = 17$ and MobileNet model have been considered in the forthcoming experiments.

2.4 Scenario 2: Comparing our method to the state of the art.

In this second scenario, our proposed method has been compared to a wide range of AIA methods in literature. For the sake of clarity, these methods have been categorized into: region-based and holistic-based, each of which contains CNN- and handcrafted-based features. It is worth noting that some works in literature use the full set of dataset's annotations (e.g., 374 concepts for Corel5K), whereas some others pick only a subset of 260 concepts. In our experiments however, we engaged both two scenarios: 374 and 260 concepts. One must know that a good AIA system should achieve an equivalence in the proportion of correctly assigning different concepts. In other words, the standard deviation in correctly assigning concepts needs to be minimized. Unfortunately, we were not able to find statistics, such as standard deviations and medians ,about the obtained results in most of the related works to compare with.

Corel-5K has had the major share of experiments for AIA tasks. Since there are many related works for which there is no room to mention here, we have involved the more recent ones in our comparison (were proposed after 2015) . Table VI.2. presents results obtained from our method compared to those of the related works using Corel-5K dataset.

Table VI.2: A comparison between our method and other recent related works in terms of Precision (P), Recall(R), F1 and N+. The involved works adopt one of the scenarios : considering 260 concepts or 374 concepts, as shown in column (N° Cpt).

	Method	N° Cpt	P	R	F1	N+
Holistic approach	CNN-R(2015)(Murthy et al., 2015)	374	32	41.3	36.1	166
	KCCA(2015) (Murthy et al., 2015)	374	39	53	44.9	184
	CCA-KNN(2015) (Murthy et al., 2015)	374	42	52	46.5	201
	Group Sparsity(2015)(X. Zhang & Liu, 2015)	260	30	33	31.4	146
	GLKNN(2015)(Su & Xue, 2015)	260	36	47	40.8	184
	MIAPS(2015)(Amiri & Jamzad, 2015)	260	39.98	42.66	41.28	177

MVSAE(2015)(R et al., 2015)	260	37	47	42	175
LJNMF(2015)(Rad & Jamzad, 2015)	260	35	43	39.1	175
SLED(2015)(X. Cao et al., 2015)	260	35	51	41.5	-
AWD-IKNN(2016)(J. Li & Yuan, 2016)	260	42	55	47.7	198
CNN-AT(2016)(Le, 2016)	374	26	17	21	88
NSIDML(2016)(R et al., 2016)	260	44.12	51.76	47.76	194
MLDL(2016)(Jing et al., 2016)	260	45	49	47	198
LDMKL(2017)(Jiu & Sahbi, 2017)	200	29	44	35	179
SDMKL(2017) (Jiu & Sahbi, 2017)	200	25	38		158
L-ADA(2017)(Ke et al., 2017)	260	31	38	34	164
NL-ADA(2017) (Ke et al., 2017)	260	32	40	36	173
MVG-NMF(2017)(R et al., 2017)	260	44	47.5	45.6	197
PRM(2017)(Khatchatoorian, 2017)	260	40.78	53.64	46.33	205
VSE-2PKNN-ML(2018)(W. Zhang et al., 2018)	260	41	52	46	205
PRM DEEP(2018)(Khatchatoorian, 2018)	260	45.3	51.73	48.3	201
CCA-KNN(2018)(Wang, X.L.; Hongwei, G.E.; Liang, 2018)	260	41	43	42	185
IDFRW(2018)(Ning et al., 2018)	260	38	49	43	185
CDNI(2018)_(Maihami & Yaghmaee, 2018)	260	29.8	32.1	30.9	162
OPSL(2018)(Xue et al., 2018)	260	38.3	55	45.2	
E2E-DCNN(2019)_(Ke et al., 2019)	260	41	55	47	192
SEM(2019)_(Ma et al., 2019)	260	37	52	43	-
L-Global CA(2019)_(Jiu & Sahbi, 2019)	260	36	45		189
S-Global CA(2019)_(Jiu & Sahbi, 2019)	260	36	46		194

	L-Classwise CA(2019) _(Jiu & Sahbi, 2019)	260	36	45		192
	LL-PLSA(2020) _(H. Song et al., 2020)	260	37	48	42	-
	RDPGKNN(2020) _(S. Chen et al., 2020)	260	40	45	40	195
	Weight-KNN(2020) _(Ma et al., 2020)	260	22	15	18	-
	khatchatoorian et al. (2020) (Khatchatoorian & Jamzad, 2020)	260	55.46	56.55	56	212
	GCN(2020) _(Z. Zhu & Hangchi, 2020)	260	48	52	49	200
	CNN-THOP(2020) _(J. Cao et al., 2020)	260	52.7	58.3	55.3	-
	SSGL(2020) _(Z. Chen et al., 2020)	260	34	47	40	190
	Zhang et al.(2020) _(W. Zhang et al., 2020)	374	60	68	64	228
	PLSA-MB(2020) _(D. Tian & Shi, 2020)	260	26	30	27.9	
	TAIA(2020) _(Ge et al., 2020)	260	38.4	48.6	42.9	177
	Y.chen et al.(2021) _(Y. Chen et al., 2021)	260	26.93	41.43	32.64	161
	TSEM(2021) _(Wei et al., 2021)	260	38	46	42	-
	TSEM+LQP(2021) _(Wei et al., 2021)	260	45	40	43	-
	SSL-AWF(Z. Li, Lin, Zhang, Ma, et al., 2021)	260	51	48	49.5	203
	CNN-SPP(Z. Li, Lin, Zhang, Ma, et al., 2021)	260	46	43	44.4	196
	HMAA (J. Chen et al., 2021)	260	43	54	48	
	MVRSC _(Zamiri & Sadoghi Yazdi, 2021)	260	54.3	42.9	47.9	
	LDA-ECC(Z. Li, Lin, Zhang, Ma, et al., 2021)	260	35	36	35.5	148
Region-based approach	MLSIA _(J. Zhang et al., 2015)	374	23.35	26.24	23.54	-
	ANNOR-G(Kuric, 2016)	260	22	29	25	129
	Zhang et al.(J. Zhang et al., 2016)	374	57.61	53.04	53.85	-
	BG(J. Zhang et al., 2019)	374	33	41		170

TG_(J. Zhang et al., 2019)	374	36	45		189
Vatani et al._(Vatani et al., 2020)	260	28	96	43	-
Our method	374	48.63	64.94	54.85	236
	260	59.45	65.01	58.89	212

From Table VI.2, it evidently appears that our proposed segmentation based AIA method outperforms the majority of the stated related works in both scenarios of 274 and 260 concepts. If we take as instance the top two F1 scores yielded by the related works Khatchatoorian et al. (Khatchatoorian & Jamzad, 2020) and CNN-THOP (J. Cao et al., 2020) in the scenario of 260 concepts, we can clearly see that the outcomes of our method exceeds those of both methods by 5% at least. Furthermore, the F1 score obtained by our method is at least 10% higher than that obtained by other recent studies such as GCN(2020) (Z. Zhu & Hangchi, 2020), SSL-AWF(2021)(Z. Li, Lin, Zhang, & Key, 2021), MVRSC (Kuric, 2016) , and so on. Now, if we look at the scenario of 374 concepts, we can see that our proposed method has surpassed all other methods except for the Vatani et al. (Vatani et al., 2020). However, if we compare the latter in terms of N+, we can see that our method outperforms it by 8 concepts. This means that our method is capable of appropriately assigning 8 more concepts than the latter. And, as previously said, it is not sufficient for a technique to achieve high accuracy alone; but should also acquire the meaning of the greatest number possible of concepts.

To further analyze the outcomes of our method, we have calculated statistics of P, R and F1 and presented them using a box plot. Figure VI.4 resumes some statistics about how our proposed method learns the meaning of concepts.

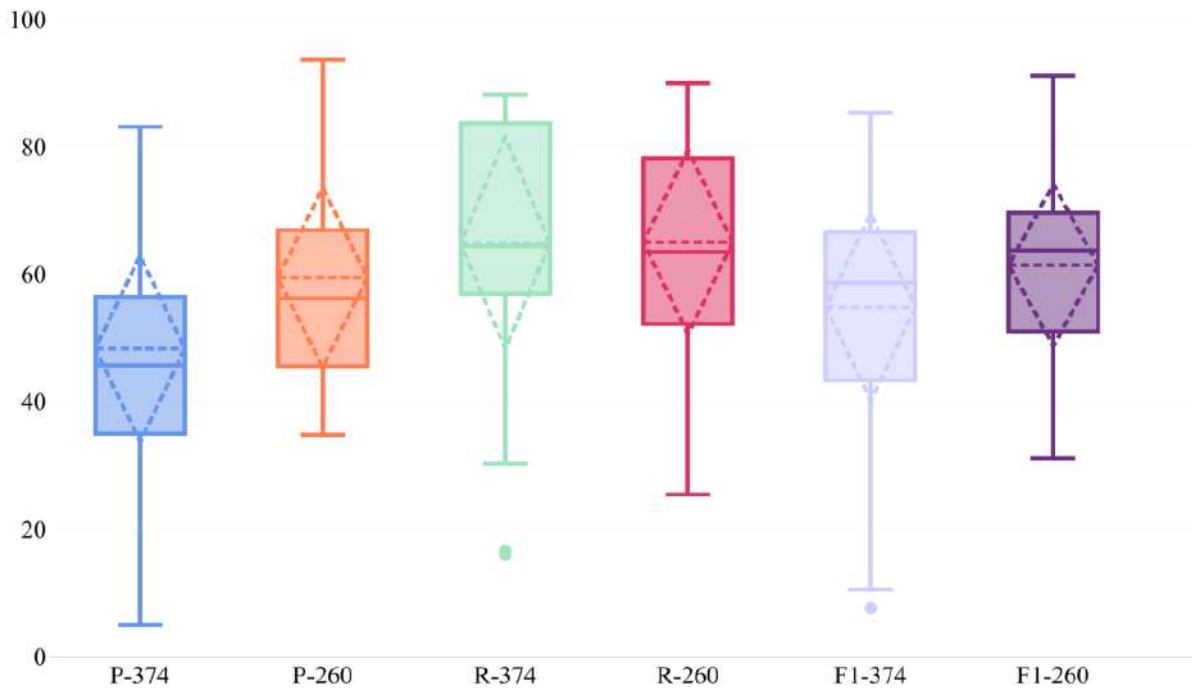


Figure VI.4: statistical description of how our proposed method learns the meanings of concepts accurately and in a balanced manner. Precision and recall are denoted by the letters P and R, respectively, and the following number denotes the number of concepts utilized in the experiment.

From the first glance, it appears that there is a compromise between precision and recall based on the used number of concepts. With 374 concepts, for instance, our system achieved a recall that's far higher than the precision. When it comes to 260 concepts, however, the precision remarkably improved whereas the recall slightly decreased. As the depicted standard deviation (≈ 14 in both cases) indicates, our proposed technique aids in the balanced learning of various concepts. With a median of 45.7 in the scenario of 374 concepts, our findings indicate that more than half of the images were annotated with at least two to three accurate concepts, which is a significant number given the large number of concepts (374 concepts). Nonetheless, the number of correctly annotated images with two to three concepts increases substantially in case of 260 images, resulting in 75% rate. It should be noted that manually annotating images involves some subjectivity or mistakes, which results in the appearance of certain outliers, as seen in Figure VI.4.

On one hand, the approach proposed in the work Zhang et al. (J. Zhang et al., 2016) relies totally on finding the semantic relatedness among pre-segmented regions based on a wide

range of handcrafted features (Aiadi et al., 2019; Kaoudja et al., 2019). By understanding the logic that connects different concepts, the system became able to learn concepts regardless of their narrow use. On the other hand, the idea in the work of khatchatoorian et al. (Khatchatoorian & Jamzad, 2020) revolves around employing the CNN as a black box and letting it learn everything by itself. However, we have taken advantage of both the methods by applying a CNN to get a rich set of features representing the concepts and employing knn regression to understand how these concepts are related. By doing so, we have exceeded the performance of both previous techniques.

MSRC v2 dataset has also been used to assist the performance of AIA systems in various literature works, in particular those based on regions. We have conducted a comparison against some recent works on the same dataset using the same 22 concepts scenario. Due to the limited number of annotations (22 concepts only), the metric N+ has been disregarded in this comparison since it always produces the perfect result (i.e., $N+=22$). Figure VI.5 presents F1 in terms of precision and recall using the MSRC v2 dataset.

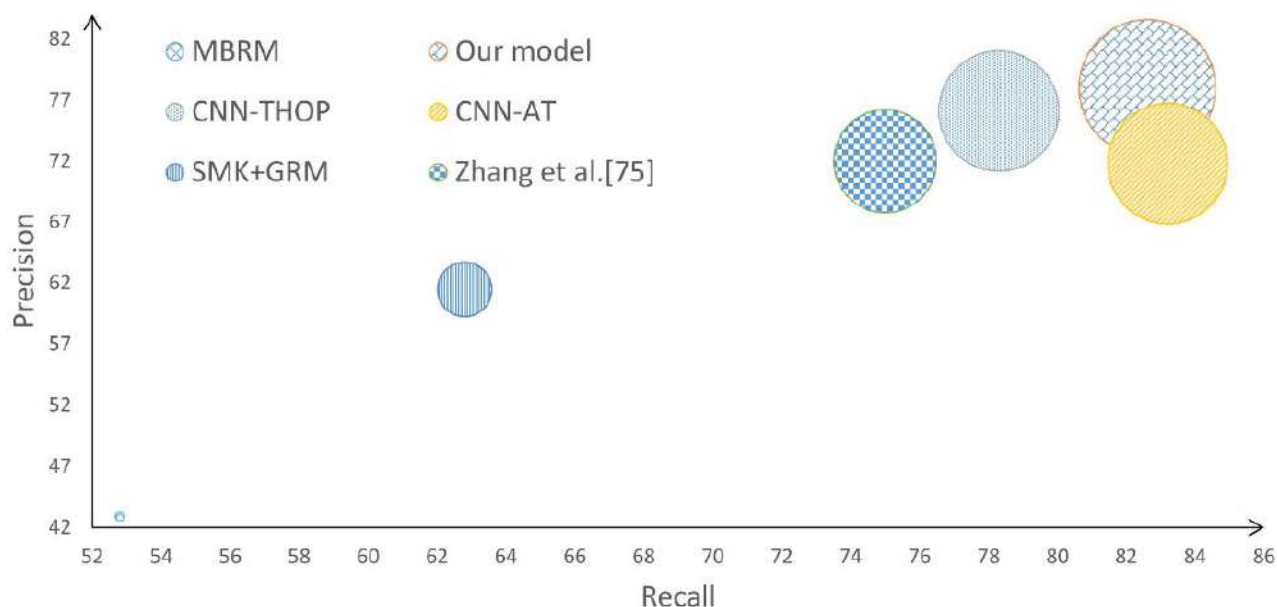


Figure VI.5: A blob chart of F1 in terms of precision and recall. The experiments were conducted on MSRC dataset, with 22 concepts, between MBRM (S. L. Feng et al., 2004), CNN-THOP (J. Cao et al., 2020), SMK+GRM (Lu & Ip, 2009), CNN-AT (Le, 2016) and the Zhang et al. (W. Zhang et al., 2020) on one hand, and our proposed method on the other hand.

Figure VI.5 clearly shows that our proposed method outperforms the others by yielding precision = 78.01% and recall = 82.6% which produce the highest F1 score of 80.24%.

However, assessing the method's performance based on a sample mean of precisions is, in many times, deceptive. Therefore, it is a common practice in AIA performance assessment procedure to evaluate the performance on each concept individually. Figure VI.6 presents a precision heatmap yielded by our method compared to the others.

	MBRM	SSK+CBKP	CNN-AT	CNN-ECC	E2E-DCNN	Zhang et al.[75]	CNN-THOP	Our method
grass	0.73	0.76	0.82	0.85	0.87	0.9	0.93	0.98
cow	0.76	0.81	0.88	0.89	0.89	0.9	0.92	1
tree	0.74	0.79	0.84	0.85	0.87	0.95	1	1
sky	0.72	0.8	0.86	0.87	0.89	0.85	0.85	0.93
building	0.73	0.8	0.85	0.86	0.89	0.96	0.96	0.97
aeroplane	0.7	0.77	0.83	0.84	0.85	0.9	0.92	0.99
mountain	0.67	0.73	0.77	0.8	0.8	0.82	0.75	0.86
face	0.85	0.91	0.96	0.95	0.97	0.9	0.89	0.93
body	0.72	0.76	0.91	0.91	0.94	0.9	0.86	0.9
car	0.81	0.86	0.9	0.91	0.94	0.95	1	1
bike	0.76	0.8	0.86	0.86	0.89	0.94	1	0.99
sheep	0.72	0.77	0.82	0.83	0.87	0.87	0.89	0.91
flower	0.74	0.79	0.86	0.85	0.9	0.9	0.83	0.86
sign	0.69	0.76	0.81	0.83	0.86	0.97	1	1
bird	0.63	0.69	0.79	0.8	0.83	0.9	0.67	0.74
water	0.74	0.76	0.81	0.82	0.86	0.85	0.86	0.88
book	0.71	0.75	0.82	0.83	0.86	0.86	0.86	0.91
chair	0.66	0.72	0.76	0.81	0.79	0.78	0.8	0.87
cat	0.77	0.84	0.9	0.89	0.92	0.9	0.8	0.84
dog	0.76	0.83	0.9	0.86	0.94	0.87	0.44	0.78
road	0.76	0.8	0.86	0.89	0.88	0.89	0.93	0.97
boat	0.69	0.72	0.77	0.81	0.83	0.85	0.88	0.91

Figure VI .6precision heatmap generated from the precision per concept produced by each method. Lower precisions are indicated by darker cells, and vice versa. The methods involved in this experiment are MBRM, SSK-CBKP, CNN-AT,CNN-ECC, E2E-DCNN(2019),CNN-THOP, Zhang et al. and our method.

As it appears from Figure VI.6, CNN-THOP and Our method have outperformed the others by yielding perfect precisions with four concepts. Furthermore, our method has achieved more than 0.98 for another 3 concepts, namely : grass, aeroplane and bike. If we take the third quantile for both methods (≈ 0.93 for CNN-THOP and ≈ 0.99 for our method) as an instance, we can deduce that the concepts that have been appropriately grasped are far more than CNN-THOP. Furthermore, our approach has a standard deviation of 0.7, whereas CCN-THOP has a standard deviation of 0.14, indicating that the former has a better balance in

learning concepts, whilst the latter only concentrates on a few of them. The outcomes of this experiment prove that guiding a CNN-based AIA system through a preprocess of image segmentation could highly improve the results.

A poor performance of an AIA system doesn't always reflect inefficiency; in many cases, it is a result of poorly annotated dataset. To further clarify this last argument, we have collected in list some images in which the ground truth doesn't accurately reflect the content of the image. Table VI.3 shows a list of test images with their respective ground truth and annotations given by AIA systems.

Table VI.3. A list of images with their respective ground truths and given annotations. Concepts in bold indicate that they are parts of the ground truth.






	Groundtruth	CMRM	Our method
	car, tracks, grass	water, tree, sky, people, grass	car, tracks , turn, prototype
	sky, tree, castle	people, building, oahu, water, tree	sky , clouds, tree , house
	flowers, petals, leaf	sky, water, people, tree, grass	leaf, flowers, petals , stems
	flowers, tree, sky	flowers, tree , grass, lawn, sky	sky, tree, flower , tulip
	sky, plane, runway	plane, jet, sky , cars, tracks	plane, runway , prop

Table VI.3 shows that, compared to the ground truth, some annotations have been indeed assigned, some replaced with their synonyms and some others have been completely omitted. If we take the image number 3 as an instance, we can see that the precision of the annotation process is 50% (i.e., two out of three concepts from the ground truth have been assigned to the image by the AIA). However, a careful inspection reveals that all the assigned concepts do indeed describe the image (image 2 contains clouds and a house). The same goes for the rest of the images.

.5 Scenario 3: Computing cost

When an algorithm is dedicated to being utilized with entities with restricted sources of power or poor processing capacity, its speed is an essential factor in determining its performance. In this experiment, we evaluate and compare our method to other common AIA methods in terms of time consumed in the annotation process. Table 4 shows the result of comparing our method to other famous methods in terms of time consumption during annotation.

Table VI.4. Time consumed, in seconds, for annotating one image with five concepts.

	Our method	SKL-CRM	MLDL	2PKNN	tagProp
consumed time	1.2	27	24.6	0.6	0.6

From Table VI.4, it appears that our method has a relatively acceptable time for annotating images. This can be attributed to the sample scheme we adopt that doesn't require complicated calculations such as the case with MLDL and SKL-CRM. This is because the present method places a strong emphasis on speed and minimal computation, which can be proved by the used sample region growing algorithm JSEG for image segmentation and off-the-shelf features extracted from the fastest network MobileNet that is dedicated for mobiles. The pretrained CNN is employed in a manner that doesn't require any further training or finetuning, which reduces the amount of computing needed. These criteria grant rapidity and low consumption of resources and make our method suitable for mobiles or other small entities.

3. Conclusion

In this chapter, we presented a model for the image annotation extension. This model has the advantage of being generalized to different types of images. It is defined by a mixture of Cnn and knn regression for which we have combined visual and textual characteristics. Experimental results on Corel-5k, and MSRC are competitive, while maintaining both high precision and high recall in a balanced way. The model has also been used to reduce silence.

Chapter VII: General conclusion

1. conclusion	10100
2. Perspectives	10102

1. conclusion

Automatic image annotation is a difficult subject to solve since it incorporates real-world photographs with various labels. It's also a fundamental problem in computer vision with a wide range of applications, and it can help with other visual learning tasks like image captioning and scene recognition, among others. We started this thesis by describing the automatic picture annotation problem, convincing the reader to care about it, and outlining the challenges. Then we looked at it from two different angles. We made some essential dataset and evaluation metrics related observations in Chapter 2 by comprehensive empirical study, which are important to create systemic breakthroughs in the picture annotation region. We concluded that per-label metrics are a stronger indicator of an annotation method's performance than per-image metrics. We demonstrated the absence of diversity in existing picture annotation datasets using the proposed measures. We've highlighted the points to keep in mind when creating fresh datasets and methodologies for the picture annotation assignment in the future. Then, in Chapters 3-4 we have studied the visual content-based images annotation techniques, in particular image segmentation, features extraction, and machine/deep learning. There are also comparisons between the many algorithms that have been demonstrated. Following that, many Machine-Learning algorithms were developed to automate the annotation process. The main challenge in machine learning is that a poor data representation reduces the quality of the produced results and results in worse performance when compared to a good data representation. As a result, feature engineering has long been regarded as an essential study field in machine learning. It concentrates on getting more detailed information from raw data. Multiple research investigations have resulted as a result of this. Deep-Learning algorithms (DL) essentially allow for feature extraction to be done automatically. This enables researchers to extract discriminative features even with limited domain knowledge, reducing human effort. Furthermore, recent development in this area demonstrates that deep learning algorithms, particularly CNN, can solve the annotation challenge. However, there are still a number of unresolved difficulties in the areas of object detection and image captioning that need to be addressed in future research. The methods we used to conduct our studies also enable new possibilities to analyze numerous areas of research. In chapters 5-6, we explained the new method used in automatic image annotation in

detail and compared it with the various works in this field with an analytical study on the data and results that prove the quality of our method.

In this thesis, we introduced an automatic image annotation system in which segmentation JSEG algorithm, a convolutional neural network named MobileNet, and KNN regression methods have been employed. MobileNet has been adopted to grant a rich representation of regions generated by JSEG, and KNN regressor is employed to understand how these concepts are related. After tuning the best values of our method, it has been compared against other methods in terms of precision, recall, F1, N+, and computing time. The two common scenarios of 374 and 260 concepts have been taken into account for the dataset Corel-5K. F1 of 54.85% and N+ of 236 for the first scenario and F1 of 58.89% and N+ of 212 for the second scenario have been achieved. These results indicate the superiority of the proposed approach compared to a wide range of related works. Furthermore, a statistical analysis has been carried out on the outcoming of our method and has proved that our proposed method aids in more balanced learning of different concepts. To further prove the superiority of our method, it has been compared against other region-based works on the MSRC v2 dataset. Results proved that the concepts corresponding to the third quartile achieve more than 99% precision, which is an important amount of concepts. Since the present method places a strong emphasis on speed and minimal computation, we compared it against other common methods in terms of time consumption. Results proved its rapidity and low consumption of resources which make it suitable for mobiles or other small entities. The experiments also demonstrated that the precision yielded by our method is somewhat biased due to the poor quality of the ground truth. Therefore, our method should be exploited in enhancing the ground truth of manually annotated datasets.

The thesis is a contribution to the field of automatic image annotation which will improve the annotation performance by accurate tagging of pictures. The automation in the annotation is improved by understanding the image contents. Segmentation of images using proper techniques leads to better region finding and improves annotation performance. The features extracted at a low level must constitute the desired concept at a high level. This will bridge the gap from low to high-level features. This has been proven by annotation accuracy with remarkable improvement. The experiments carried out throughout the research work suggests

segmentation algorithms, features with their extraction mechanism, and machine learning techniques for improvement of annotation accuracy performance

2. Perspectives

Image annotation is one of the most challenging research areas in computer vision, but with our model, we have achieved satisfactory results. From the perspectives of our work, other investigations can be carried out to improve the proposed method, namely:

In the short term:

- ✓ The proposed automatic image annotation system architecture has improved the accuracy and efficiency of tagging and retrieval. A high precision reflects the created tags' reliability, enhancing search results.

Medium-term :

- ✓ Features are the inputs for machine learning techniques that represent high-level concepts at a low level. More, the gap between these levels generates poor results. We plan to enrich our system by using other visual characteristics for hearing the collection of usable descriptors.
- ✓ In the system developed, the region generation done by using segmentation techniques plays a vital role in understanding parts of a picture. Improvement in segmentation algorithms is still possible and can increase the overall performance of tagging.

Long-term:

- ✓ Recognizing and detecting object leads to finding and defining concepts that represent the image well. We are expanding this framework to high accuracy in object recognition and object detection.

Bibliography

- Adnan, M. M., Mohd-Rahim, M. S., Khaleel, S. M., & Al-Jawaheri, K. (2019). A Survey Automatic Image Annotation Based on Machine Learning Models. *Journal of Engineering and Applied Sciences*, 14(20), 7627–7635. <https://doi.org/10.36478/jeasci.2019.7627.7635>
- Agarwal, S. (2014). Data mining: Data mining concepts and techniques. *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*, 203–207. <https://doi.org/10.1109/ICMIRA.2013.45>
- Ahn, L. Von, & Dabbish, L. (2004). *Labeling Images with a Computer Game*. 6(1), 319–326.
- Aiadi, O., Kherfi, M. L., & Khaldi, B. (2019). Automatic date fruit recognition using outlier detection techniques and Gaussian mixture models. *Electronic Letters on Computer Vision and Image Analysis*, 18(1), 52–75. <https://doi.org/10.5565/rev/elcvia.1041>
- Aloun, M. S., Hitam, M. S., Yussof, W. N. J. H. W., Abdul Hamid, A. A. K., & Bachok, Z. (2019). Modified JSEG algorithm for reducing over-segmentation problems in underwater coral reef images. *International Journal of Electrical and Computer Engineering*, 9(6), 5244–5252. <https://doi.org/10.11591/ijece.v9i6.pp5244-5252>
- Amiri, S. H., & Jamzad, M. (2015). Efficient multi-modal fusion on supergraph for scalable image annotation. *Pattern Recognition*, 1–13. <https://doi.org/10.1016/j.patcog.2015.01.015>
- Analysis, V. I. (2006). *A CONTENT-BASED IMAGE RETRIEVAL SCHEME IN JPEG COMPRESSED DOMAIN* Zhe-Ming Lu 1 , 2 , Su-Zhi Li 2 and Hans Burkhardt 1. 2(4), 831–839.
- Arockiam, L., Baskar, S. S., & Jeyasimman, L. (2012). Clustering methods and algorithms in data mining: Review. *Asian Journal of Information Technology*, 11(1), 40–44. <https://doi.org/10.3923/ajit.2012.40.44>
- Avenue, F. (n.d.). *TRADEMARK RETRIEVAL USING CONTOUR-SKELETON STROKE CLASSIFICATION*. c, 1–4.
- B.S. Manjunath, W. Y. M. (1996). Texture features for browsing and retrieval of image data. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 18(8), 837–842.
- Bakliwal, P., & Jawahar, C. V. (2016). Active learning based image annotation. *2015 5th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, NCVPRIPG 2015*. <https://doi.org/10.1109/NCVPRIPG.2015.7490061>
- Ballan, L., Uricchio, T., Seidenari, L., & Del Bimbo, A. (2014). A cross-media model for automatic image annotation. *ICMR 2014 - Proceedings of the ACM International Conference on*

Bibliography

- Multimedia Retrieval 2014*, 1(Micc), 73–80. <https://doi.org/10.1145/2578726.2578728>
- Bao, B. K., Li, T., & Yan, S. (2012). Hidden-concept driven multilabel image annotation and label ranking. *IEEE Transactions on Multimedia*, 14(1), 199–210. <https://doi.org/10.1109/TMM.2011.2170557>
- Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D. M., & Jordan, M. I. (2003). Matching Words and Pictures. *Journal of Machine Learning Research*, 3(6), 1107–1135. <https://doi.org/10.1162/153244303322533214>
- Bay, H., Tuytelaars, T., & Gool, L. Van. (2006). SURF: Speeded Up Robust Features. 404–417.
- Ben, H., Pan, Y., Li, Y., Yao, T., Hong, R., Wang, M., & Mei, T. (2021). Unpaired Image Captioning with Semantic-Constrained Self-Learning. *IEEE Transactions on Multimedia*, 9210(c), 1–13. <https://doi.org/10.1109/TMM.2021.3060948>
- Ben Rejeb, I., Ouni, S., Barhoumi, W., & Zagrouba, E. (2018). Fuzzy VA-Files for multi-label image annotation based on visual content of regions. *Signal, Image and Video Processing*, 12(5), 877–884. <https://doi.org/10.1007/s11760-017-1233-1>
- Berens, J., Finlayson, G. D., & Qiu, G. (2000). Image indexing using compressed colour histograms. *IEE Proceedings - Vision, Image, and Signal Processing*, 147(4), 349. <https://doi.org/10.1049/ip-vis:20000630>
- Bhagat, P. K., & Choudhary, P. (2018). Image annotation: Then and now. *Image and Vision Computing*, 80, 1–23. <https://doi.org/10.1016/j.imavis.2018.09.017>
- Bhatt, H. S., Bharadwaj, S., Singh, R., & Vatsa, M. (2010). On matching sketches with digital face images. *IEEE 4th International Conference on Biometrics: Theory, Applications and Systems, BTAS 2010*. <https://doi.org/10.1109/BTAS.2010.5634507>
- Blei, D. M., & Jordan, M. I. (2003). *Modeling annotated data*. 127. <https://doi.org/10.1145/860458.860460>
- Borràs, A., Tous, F., Lladós, J., & Vanrell, M. (2003). High-level clothes description based on colour-texture and structural features. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2652, 108–116. https://doi.org/10.1007/978-3-540-44871-6_13
- Bouchakwa, M., Ayadi, Y., & Amous, I. (2020). A review on visual content-based and users' tags-based image annotation: methods and techniques. *Multimedia Tools and Applications*, 79(29–30), 21679–21741. <https://doi.org/10.1007/s11042-020-08862-1>
- BOUZAYANI, A. (2018). *Extension automatique de l'annotation d'images pour la recherche et la classificatio*.

Bibliography

- Bratasanu, D., Nedelcu, I., & Datcu, M. (2011). Bridging the Semantic Gap for Satellite Image Annotation and Automatic Mapping Applications. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4(1), 193–204.
<https://doi.org/10.1109/JSTARS.2010.2081349>
- C., G. R., & E., W. R. (2002). *Digital image processing*.
- C., Y., M., D., & F., F. (2005). *Image content annotation using Bayesian framework and complement components analysis*. <https://ieeexplore.ieee.org/abstract/document/1529970>
- Cao, J., Zhao, A., & Zhang, Z. (2020). Automatic image annotation method based on a convolutional neural network with threshold optimization. *PLoS ONE*, 15(9 September), 1–21. <https://doi.org/10.1371/journal.pone.0238956>
- Cao, X., Member, S., Zhang, H., Guo, X., Liu, S., Meng, D., & Member, S. (2015). *SLED: Semantic Label Embedding Dictionary Representation for Multilabel Image Annotation*. 24(9), 2746–2759.
- Carbonell, J. G. (1981). Machine learning research. *ACM SIGART Bulletin*, 18(77), 29–29.
<https://doi.org/10.1145/1056743.1056744>
- Carbonetto, P., de Freitas, N., & Barnard, K. (2004). A Statistical Model for General Contextual Object Recognition. In *Statistics* (pp. 350–362). https://doi.org/10.1007/978-3-540-24670-1_27
- Carneiro, G., Chan, A. B., Moreno, P. J., & Vasconcelos, N. (2007). Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3), 394–410. <https://doi.org/10.1109/TPAMI.2007.61>
- Carneiro, G., & Vasconcelos, N. (2005). Formulating semantic image annotation as a supervised learning problem. *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, II*, 163–168. <https://doi.org/10.1109/CVPR.2005.164>
- Carson, C., Thomas, M., Belongie, S., Hellerstein, J. M., & Malik, J. (1999). Blobworld: A system for region-based image indexing and retrieval. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1614, 509–517.
https://doi.org/10.1007/3-540-48762-x_63
- Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, 288–296.
- Chang, S.-K., & Jungert, E. (n.d.). A spatial knowledge structure for image information systems using symbolic projections. *ACM '86: Proceedings of 1986 ACM Fall Joint Computer Conference*,

Bibliography

- 79–86. <https://dl.acm.org/doi/abs/10.5555/324493.324548>
- Chen, B., Li, J., Lu, G., Yu, H., & Zhang, D. (2020). Label co-occurrence learning with graph convolutional networks for multi-label chest X-ray image classification. *IEEE Journal of Biomedical and Health Informatics*, 24(8), 2292–2302.
<https://doi.org/10.1109/JBHI.2020.2967084>
- Chen, G., Zhang, J., Wang, F., & Zhang, C. (2009). *Efficient Multi-label Classification with Hypergraph Regularization*. 1658–1665.
- Chen, J., Ying, P., Fu, X., Luo, X., Guan, H., & Wei, K. (2021). Automatic tagging by leveraging visual and annotated features in social media. *IEEE Transactions on Multimedia*, 9210(c), 1–12.
<https://doi.org/10.1109/TMM.2021.3055037>
- Chen, S., Wang, M., & Chen, X. (2020). Image annotation via reconstitution graph learning model. *Wireless Communications and Mobile Computing*, 2020.
<https://doi.org/10.1155/2020/8818616>
- Chen, X., Mu, Y., Yan, S., & Chua, T. (2010). Efficient Large-Scale Image Annotation by Probabilistic Collaborative Multi-Label Propagation Categories and Subject Descriptors. *MM '10: Proceedings of the 18th ACM International Conference on Multimedia*, 35–44.
<https://doi.org/https://doi.org/10.1145/1873951.1873959>
- Chen, Y., Liu, L., Tao, J., Chen, X., Xia, R., Zhang, Q., Xiong, J., Yang, K., & Xie, J. (2021). The image annotation algorithm using convolutional features from intermediate layer of deep learning. *Multimedia Tools and Applications*, 80(3), 4237–4261.
<https://doi.org/10.1007/s11042-020-09887-2>
- Chen, Z., Wang, M., Gao, J., & Li, P. (2020). Image Annotation based on Semantic Structure and Graph Learning. *Proceedings - IEEE 18th International Conference on Dependable, Autonomic and Secure Computing, IEEE 18th International Conference on Pervasive Intelligence and Computing, IEEE 6th International Conference on Cloud and Big Data Computing and IEEE 5th Cyber*, 451–456.
<https://doi.org/10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00085>
- Cheng, Q., Zhang, Q., Fu, P., Tu, C., & Li, S. (2018). A survey and analysis on automatic image annotation. *Pattern Recognition*, 79, 242–259. <https://doi.org/10.1016/j.patcog.2018.02.017>
- Chua, T., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). *NUS-WIDE : A Real-World Web Image Database from National University of Singapore*. 0–8.
- Colleges, G. T. U. A., Academy, O., Academy, O., Academy, O., Science, A. C., Technology, I., & Science, A. C. (2014). Microsoft COCO. *Eccv, June*, 740–755.
- Cross, G. R. (1983). *Markov Random Field Texture Models*. 1, 25–39.

Bibliography

- Cusano, C., Ciocca, G., & Schettini, R. (2003). *Image annotation using SVM* (S. Santini & R. Schettini (eds.); pp. 330–338). <https://doi.org/10.1117/12.526746>
- D., F. J., D., V. F., Van, D. A., K., F. S., & F., H. J. (1996). *Computer Graphics: Principles and Practice*. javascript:void(0);
- Darwish, S. M. (2016). *Combining firefly algorithm and Bayesian classifier : new direction for automatic multilabel image annotation*. 763–772. <https://doi.org/10.1049/iet-ipr.2015.0492>
- David R. Hardoon, J. S.-T. (2003). KCCA FOR DIFFERENT LEVEL PRECISION IN CONTENT-BASED IMAGE RETRIEVAL. *Science*, 0–5.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm . *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–22. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>
- Deng, Y., Manjunath, B. S., Kenney, C., Moore, M. S., & Shin, H. (2001). An efficient color representation for image retrieval. *IEEE Transactions on Image Processing*, 10(1), 140–147. <https://doi.org/10.1109/83.892450>
- Dhanachandra, N., Manglem, K., & Chanu, Y. J. (2015). Image Segmentation Using K-means Clustering Algorithm and Subtractive Clustering Algorithm. *Procedia Computer Science*, 54, 764–771. <https://doi.org/10.1016/j.procs.2015.06.090>
- Dietterich, T. G. (2002). Machine learning for sequential data: A review. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2396, 15–30. https://doi.org/10.1007/3-540-70659-3_2
- Dietterich, T. G., & Oregon. (1996). Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Models. *Oncogene*, 12(2), pp 1-15(265-275).
- Ding, X., Li, B., Xiong, W., Guo, W., Hu, W., Wang, B., Technology, S., & Technology, I. (2016). *Multi - Instance Multi - label Learning Combining Hierarchical Context and Its Application to Image Annotation*. 9210(MLL). <https://doi.org/10.1109/TMM.2016.2572000>
- Doggaz, N., & Ferjani, I. (2011). Image segmentation using normalized cuts and efficient graph-based segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6979 LNCS(PART 2), 229–240. https://doi.org/10.1007/978-3-642-24088-1_24
- Dutta, A. (2019). *Blending the Past and Present of Automatic Image Annotation*. <https://www.semanticscholar.org/paper/Blending-the-Past-and-Present-of-Automatic-Image-Dutta-Sivaswamy/5ed526103c5876741a962561a6e3196121ffa6e7#related-papers>
- Duygulu, P., Barnard, K., de Freitas, J. F. G., & Forsyth, D. A. (2002). Object recognition as

Bibliography

- machine translation: Learning a lexicon for a fixed image vocabulary. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2353, 97–112. https://doi.org/10.1007/3-540-47979-1_7
- Eddy, S. (2004). What is a hidden Markov model? bioinformatics. *Nature Biotechnology*, 1–5. https://scholar.google.com/citations?view_op=view_citation&continue=/scholar%3Fhl%3Den%26as_sdt%3D0,14%26scilib%3D1&citilm=1&citation_for_view=NzNAerUAAAAJ:_FxGoFyzp5QC&hl=en&oi=p
- F., L., H., Z., & D., F. D. (2003). Fundamentals of content-based image retrieval. *Multimedia Information Retrieval and Management*, 1–26. https://link.springer.com/chapter/10.1007/978-3-662-05300-3_1
- Fakhari, A., & Moghadam, A. M. E. (2013). Combination of classification and regression in decision tree for multi-labeling image annotation and retrieval. *Applied Soft Computing Journal*, 13(2), 1292–1302. <https://doi.org/10.1016/j.asoc.2012.10.019>
- Fan, J., Gao, Y., & Luo, H. (2008). Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *IEEE Transactions on Image Processing*, 17(3), 407–426. <https://doi.org/10.1109/TIP.2008.916999>
- Fan, J., Gao, Y., Luo, H., & Xu, G. (2004). Automatic image annotation by using concept-sensitive salient objects for image content representation. *Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 361–368. <https://doi.org/10.1145/1008992.1009055>
- Feng, L., & Bhanu, B. (2016). Semantic Concept Co-Occurrence Patterns for Image Annotation and Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4), 785–799. <https://doi.org/10.1109/TPAMI.2015.2469281>
- Feng, S. L., Manmatha, R., & Lavrenko, V. (2004). Multiple Bernoulli relevance models for image and video annotation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2. <https://doi.org/10.1109/cvpr.2004.1315274>
- Feng, Z., Jin, R., & Jain, A. (2013). Large-scale image annotation by efficient and robust kernel metric learning. *Proceedings of the IEEE International Conference on Computer Vision, Dml*, 1609–1616. <https://doi.org/10.1109/ICCV.2013.203>
- Fine, S., Singer, Y., & Tishby, N. (1998). The hierarchical hidden Markov model: Analysis and applications. *Machine Learning*, 32(1), 41–62. <https://doi.org/10.1023/A:1007469218079>
- Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Qian Huang, Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., & Yanker, P. (1995). Query by image and video content:

Bibliography

- the QBIC system. *Computer*, 28(9), 23–32. <https://doi.org/10.1109/2.410146>
- Frstner, W. (1994). A framework for low level feature extraction. *Computer*.
- Fu, H., Zhang, Q., & Qiu, G. (2012). Random forest for image annotation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7577 LNCS(PART 6), 86–99. https://doi.org/10.1007/978-3-642-33783-3_7
- Gårding, J., & Lindeberg, T. (1996). Direct computation of shape cues using scale-adapted spatial derivative operators. *International Journal of Computer Vision*, 17(2), 163–191. <https://doi.org/10.1007/bf00058750>
- Ge, H., Zhang, K., Hou, Y., Yu, C., Zhao, M., Wang, Z., & Sun, L. (2020). Two-stage Automatic Image Annotation Based on Latent Semantic Scene Classification. *Proceedings of the International Joint Conference on Neural Networks*. <https://doi.org/10.1109/IJCNN48605.2020.9207176>
- Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6*(6), 721–741. <https://doi.org/10.1109/TPAMI.1984.4767596>
- Ghoshal, A., Ircing, P., & Khudanpur, S. (2005). Hidden Markov models for automatic annotation and content-based retrieval of images and video. *SIGIR 2005 - Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 544–551. <https://doi.org/10.1145/1076034.1076127>
- Glotin, H., & Tollari, S. (2005). *Fast Image Auto-Annotation With Visual Vector Approximation Clusters*.
- Goh, K. S., Chang, E. Y., & Li, B. (2005). Using one-class and two-class SVMs for multiclass image annotation. *IEEE Transactions on Knowledge and Data Engineering*, 17(10), 1333–1346. <https://doi.org/10.1109/TKDE.2005.170>
- Gong, Y., Jia, Y., Leung, T., Toshev, A., & Ioffe, S. (2013). *Deep Convolutional Ranking for Multilabel Image Annotation*. 1–9. <http://arxiv.org/abs/1312.4894>
- Grangier, D., & Bengio, S. (2008). A discriminative kernel-based approach to rank images from text queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8), 1371–1384. <https://doi.org/10.1109/TPAMI.2007.70791>
- Guang-Tsai, L., W., T. R., & K., G. B. (1999). High-frequency characterization of power/ground-plane structures. *IEEE Transactions on Microwave Theory and Techniques*, 47((5)), 562–569.
- Guillaumin, M., Mensink, T., Verbeek, J., & Schmid, C. (2009). TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. *Proceedings of the IEEE*

Bibliography

- International Conference on Computer Vision*, 309–316.
<https://doi.org/10.1109/ICCV.2009.5459266>
- Hambali, H. A., Abdullah, S. L. S., Jamil, N., & Harun, H. (2017). Fruit classification using neural network model. *Journal of Telecommunication, Electronic and Computer Engineering*, 9(1–2), 43–46.
- He, X. J., Zhang, Y., Lok, T. M., & Lyu, M. R. (2006). A new feature of uniformity of image texture directions coinciding with the human eyes perception. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3614 LNAI, 727–730. https://doi.org/10.1007/11540007_90
- He, X., Zemel, R. S., & Carreira-Perpiñán, M. Á. (2004). Multiscale conditional random fields for image labeling. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2. <https://doi.org/10.1109/cvpr.2004.1315232>
- Hervé, N. (2007). *Image annotation : which approach for realistic databases ?* 0–7.
- Hong, R., Wang, M., Gao, Y., Tao, D., & Member, S. (2014). *Image Annotation By Multiple-Instance Learning With Discriminative Feature Mapping and Selection*. 44(5), 669–680.
- Hong, Z., & Jiang, Q. (2008). *Hybrid Content-based Trademark Retrieval using Region and Contour Features*. 1163–1168. <https://doi.org/10.1109/WAINA.2008.82>
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. <http://arxiv.org/abs/1704.04861>
- Hu, M., Yang, Y., Shen, F., Zhang, L., Shen, H. T., & Fellow, X. L. (2017). *Robust Web Image Annotation via Exploring Multi-facet and Structural Knowledge*. 7149(c), 1–13.
<https://doi.org/10.1109/TIP.2017.2717185>
- Huang, J., Kumar, S. R., Mitra, M., Zhu, W. J., & Zabih, R. (1997). Image indexing using color correlograms. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 762–768. <https://doi.org/10.1109/cvpr.1997.609412>
- Islam, M. M., Zhang, D., & Lu, G. (2008). A geometric method to compute directionality features for texture images. *2008 IEEE International Conference on Multimedia and Expo, ICME 2008 - Proceedings*, 3, 1521–1524. <https://doi.org/10.1109/ICME.2008.4607736>
- Islam, M. M., Zhang, D., & Lu, G. (2010). *Region Based Color Image Retrieval Using Curvelet Transform* (pp. 448–457). https://doi.org/10.1007/978-3-642-12304-7_42
- Ivasic-Kos, M., Pobar, M., & Ribaric, S. (2016). Two-tier image annotation model based on a multi-label classifier and fuzzy-knowledge representation scheme. *Pattern Recognition*, 52, 287–305. <https://doi.org/10.1016/j.patcog.2015.10.017>

Bibliography

- Jaiswal, S., & Pandey, M. K. (2021). A Review on Image Segmentation. *Advances in Intelligent Systems and Computing*, 1187, 233–240. https://doi.org/10.1007/978-981-15-6014-9_27
- Jeon, J., Lavrenko, V., & Manmatha, R. (2003). Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. *SIGIR Forum (ACM Special Interest Group on Information Retrieval), SPEC. ISS.*, 119–126. <https://doi.org/10.1145/860458.860459>
- Jing, X. Y., Wu, F., Li, Z., Hu, R., & Zhang, D. (2016). Multi-Label Dictionary Learning for Image Annotation. *IEEE Transactions on Image Processing*, 25(6), 2712–2725. <https://doi.org/10.1109/TIP.2016.2549459>
- Jiu, M., & Sahbi, H. (2017). Nonlinear Deep Kernel Learning for Image Annotation. *IEEE Transactions on Image Processing*, 26(4), 1820–1832. <https://doi.org/10.1109/TIP.2017.2666038>
- Jiu, M., & Sahbi, H. (2019). *Deep Context-Aware Kernel Networks*. 1–17. <http://arxiv.org/abs/1912.12735>
- Johnson, J. (2015). Love Thy Neighbors : Image Annotation by Exploiting Image Metadata. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Jun, T. (2010). A color image segmentation algorithm based on region growing. *ICCET 2010 - 2010 International Conference on Computer Engineering and Technology, Proceedings*, 6, 634–637. <https://doi.org/10.1109/ICCET.2010.5486012>
- Kaganami, H. G., & Beiji, Z. (2009). Region-based segmentation versus edge detection. *IIH-MSP 2009 - 2009 5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 1217–1221. <https://doi.org/10.1109/IIH-MSP.2009.13>
- Kalayeh, M. M., Idrees, H., & Shah, M. (2014). NMF-KNN: Image Annotation Using Weighted Multi-view Non-negative Matrix Factorization. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 184–191. <https://doi.org/10.1109/CVPR.2014.31>
- Kang, W. X., Yang, Q. Q., & Liang, R. P. (2009). The comparative research on image segmentation algorithms. *Proceedings of the 1st International Workshop on Education Technology and Computer Science, ETCS 2009*, 2, 703–707. <https://doi.org/10.1109/ETCS.2009.417>
- Kaoudja, Z., Kherfi, M. L., & Khaldi, B. (2019). An efficient multiple-classifier system for Arabic calligraphy style recognition. *2019 International Conference on Networking and Advanced Systems (ICNAS)*, 1–5. <https://doi.org/10.1109/ICNAS.2019.8807829>
- Ke, X., Zhou, M., Niu, Y., & Guo, W. (2017). Data equilibrium based automatic image annotation by fusing deep model and semantic propagation. *Pattern Recognition*, 71, 60–77. <https://doi.org/10.1016/j.patcog.2017.05.020>
- Ke, X., Zou, J., & Niu, Y. (2019). End-to-End Automatic Image Annotation Based on Deep

Bibliography

- CNN and Multi-Label Data Augmentation. *IEEE Transactions on Multimedia*, 21(8), 2093–2106. <https://doi.org/10.1109/TMM.2019.2895511>
- Khatchatoorian, A. G. (2017). *Post rectifying methods to improve the accuracy of image annotation*. 406–412.
- Khatchatoorian, A. G. (2018). *An Image Annotation Rectifying Method Based on Deep Features*. 88–92.
- Khatchatoorian, A. G., & Jamzad, M. (2020). Architecture to improve the accuracy of automatic image annotation systems. *IET Computer Vision*, 14(5), 214–223. <https://doi.org/10.1049/iet-cvi.2019.0500>
- Khattab, D., Ebied, H. M., Hussein, A. S., & Tolba, M. F. (2014). Color image segmentation based on different color space models using automatic GrabCut. *Scientific World Journal*, 2014. <https://doi.org/10.1155/2014/126025>
- Kodituwakku, S., & Selvarajah, S. (2004). Comparison of color features for image retrieval. *Indian Journal of Computer Science and ...*, 1(3), 207–211. <http://www.ijcse.com/docs/IJCSE10-01-03-06.pdf>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. <https://doi.org/10.1145/3065386>
- Kuric, E. (2016). *ANNOR: Efficient Image Annotation Based on Combining Local and Global Features*.
- Kuroda, K., & Hagiwara, M. (2002). An image retrieval system by impression words and specific object names-IRIS. *Neurocomputing*, 43(1–4), 259–276. [https://doi.org/10.1016/S0925-2312\(01\)00344-7](https://doi.org/10.1016/S0925-2312(01)00344-7)
- Kwasnicka, H., & Paradowski, M. (2006). On evaluation of image auto-annotation methods. *Proceedings - ISDA 2006: Sixth International Conference on Intelligent Systems Design and Applications*, 2, 353–358. <https://doi.org/10.1109/ISDA.2006.253861>
- Kyung-Wook, P., Jeong, J.-W., & Dong-Ho Lee. (2007). OLYBIA: ontology-based automatic image annotation system using semantic inference rules. *International Conference on Database Systems for Advanced Applications DASFAA 2007: Advances in Databases: Concepts, Systems and Applications*, 485–496.
- Lai, S., Zhu, Y., & Jin, L. (2020). Encoding Pathlet and SIFT Features with Bagged VLAD for Historical Writer Identification. *IEEE Transactions on Information Forensics and Security*, 15(c), 3553–3566. <https://doi.org/10.1109/TIFS.2020.2991880>
- LALAOUI, L., & MOHAMADI, T. (2013). A comparative study of Image Region-Based Segmentation Algorithms. *International Journal of Advanced Computer Science and Applications*,

Bibliography

- 4(6). <https://doi.org/10.14569/ijacsa.2013.040627>
- Lavrenko, V., Manmatha, R., & Jeon, J. (2004). A model for learning the semantics of pictures. *Advances in Neural Information Processing Systems*.
- Le, H. M. (2016). *Fully Automated Multi-label Image Annotation by Convolutional Neural Network and Adaptive Thresholding*.
- Lee, A. J. T., & Chiu, H. (2003). 2D Z-string: A new spatial knowledge representation for image databases. 24, 3015–3026. [https://doi.org/10.1016/S0167-8655\(03\)00162-4](https://doi.org/10.1016/S0167-8655(03)00162-4)
- Lee, K., & Chen, L. (2005). An efficient computation method for the texture browsing descriptor of MPEG-7 *. 23, 479–489. <https://doi.org/10.1016/j.imavis.2004.12.002>
- Lei, C., Liu, D., & Li, W. (2015). Social Diffusion Analysis with Common-Interest Model for Image Annotation. XX(XX), 1–15. <https://doi.org/10.1109/TMM.2015.2477277>
- Li, J., & Yuan, C. (2016). Automatic Image Annotation Using Adaptive Weighted Distance in Improved K Nearest Neighbors Framework. 2, 345–354. <https://doi.org/10.1007/978-3-319-48890-5>
- Li, Z., Lin, L. A. N., Zhang, C., & Key, G. (2021). A Semi-supervised Learning Approach Based on Adaptive. 17(1), 1–23.
- Li, Z., Lin, L., Zhang, C., Ma, H., Zhao, W., & Shi, Z. (2021). A Semi-supervised Learning Approach Based on Adaptive Weighted Fusion for Automatic Image Annotation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17(1), 1–23. <https://doi.org/10.1145/3426974>
- Lienhart, R., Romberg, S., & Hörster, E. (2009). Multilayer pLSA for multimodal image retrieval. *CIVR 2009 - Proceedings of the ACM International Conference on Image and Video Retrieval*, 60–67. <https://doi.org/10.1145/1646396.1646408>
- Likas, A., Vlassis, N., & Verbeek, J. (2011). The global k-means clustering algorithm Intelligent Autonomous Systems. *ISA Technical Report Series*.
- Lim, J. H., & Jin, J. S. (2005). A structured learning framework for content-based image indexing and visual query. *Multimedia Systems*, 10(4), 317–331. <https://doi.org/10.1007/s00530-004-0158-z>
- Lin, Z., Ding, G., & Hu, M. (2015). Image auto-annotation via tag-dependent random search over range-constrained visual neighbours. *Multimedia Tools and Applications*, 74(11), 4091–4116. <https://doi.org/10.1007/s11042-013-1811-3>
- Lin, Z., Ding, G., Hu, M., Wang, J., & Sun, J. (2012). Automatic image annotation using tag-related random search over visual neighbors. *ACM International Conference Proceeding Series*, 1784–1788. <https://doi.org/10.1145/2396761.2398517>

Bibliography

- Lindstaedt, S., Mörzinger, R., Sorschag, R., Pammer, V., & Thallinger, G. (2009). Automatic image annotation using visual content and folksonomies. *Multimedia Tools and Applications*, 42(1), 97–113. <https://doi.org/10.1007/s11042-008-0247-7>
- Lions, P. L., Morel, J. M., Sapiro, G., Tannenbaum, A., Witkin, P., Baudin, M., Von, V., Springer, J., Hamilton, R. S., Weldon, E. J., Tannenbaum, A., Zucker, S. W., Poggio, T. A., & Sarkar, N. (1995). *Texture Segmentation Using Fractal Dimension*. 17(1), 72–77.
- Liu, F., & Picard, R. W. (1996). *Periodicity, directionality, and randomness: World features for image modeling and retrieval*. 18(320), 722–733.
- Liu, J., Li, M., Liu, Q., Lu, H., & Ma, S. (2009). Image annotation via graph learning. *Pattern Recognition*, 42(2), 218–228. <https://doi.org/10.1016/j.patcog.2008.04.012>
- Liu, J., Wang, B., Li, M., Li, Z., Ma, W., Lu, H., & Ma, S. (2007). Dual cross-media relevance model for image annotation. *Proceedings of the ACM International Multimedia Conference and Exhibition*, 605–614. <https://doi.org/10.1145/1291233.1291380>
- Liu, Y., & Tjondronegoro, D. (2007). *A Shape Ontology Framework for Bird Classification*.
- Lomax, S., & Vadera, S. (2013). A survey of cost-sensitive decision tree induction algorithms. *ACM Computing Surveys*, 45(2). <https://doi.org/10.1145/2431211.2431215>
- Lowe, D. G. (2004). *Distinctive Image Features from Scale-Invariant Keypoints*. 1–28.
- Lu, Z., & Ip, H. H. S. (2009). Generalized relevance models for automatic image annotation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5879 LNCS, 245–255. https://doi.org/10.1007/978-3-642-10467-1_21
- Luo, J., York, N., & Savakis, A. (n.d.). *Indoor vs Outdoor Classification of Consumer Photographs Using Low-Level and Semantic Features*. 745–748.
- M. Grubinger, P. D. Clough, H. Müller, and T. D. (2006). The IAPR benchmark: A new evaluation resource for visual information systems. *In International Conference on Language Resources and Evaluation*. <http://www-i6.informatik.rwth-aachen.de/imageclef/resources/iaprtc12.tgz>
- M Saleem, R Senthilkumar, T. P. (2015). Image retrieval system by automatic annotation. *Semantic Scholar*.
- Ma, Y., Liu, Y., Xie, Q., & Li, L. (2019). CNN-feature based automatic image annotation method. *Multimedia Tools and Applications*, 78(3), 3767–3780. <https://doi.org/10.1007/s11042-018-6038-x>
- Ma, Y., Xie, Q., Liu, Y., & Xiong, S. (2020). A weighted KNN-based automatic image annotation

Bibliography

- method. *Neural Computing and Applications*, 32(11), 6559–6570.
<https://doi.org/10.1007/s00521-019-04114-y>
- Maihimi, V., & Yaghmaee, F. (2018). Automatic image annotation using community detection in neighbor images. *Physica A: Statistical Mechanics and Its Applications*, 507, 123–132.
<https://doi.org/10.1016/j.physa.2018.05.028>
- Maini, R., & Aggarwal, H. (n.d.). *Study and Comparison of Various Image Edge Detection Techniques*. 147002(3), 1–12.
- Makadia, A., Pavlovic, V., & Kumar, S. (2008). A new baseline for image annotation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5304 LNCS(PART 3), 316–329. https://doi.org/10.1007/978-3-540-88690-7_24
- Makadia, A., Pavlovic, V., & Kumar, S. (2010). Baselines for image annotation. *International Journal of Computer Vision*, 90(1), 88–105. <https://doi.org/10.1007/s11263-010-0338-6>
- Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America A*, 7(5), 923. <https://doi.org/10.1364/JOSAA.7.000923>
- Mayhew, M. B., Chen, B., & Ni, K. S. (2016). Assessing semantic information in convolutional neural network representations of images via image annotation. *Proceedings - International Conference on Image Processing, ICIP, 2016-Augus*, 2266–2270.
<https://doi.org/10.1109/ICIP.2016.7532762>
- McConville, R., Santos-Rodríguez, R., Piechocki, R. J., & Craddock, I. (2020). N2D: (not too) deep clustering via clustering the local manifold of an autoencoded embedding. *Proceedings - International Conference on Pattern Recognition*, 5145–5152.
<https://doi.org/10.1109/ICPR48806.2021.9413131>
- McGregor, A., Hall, M., Lorier, P., & Brunskill, J. (2004). Flow clustering using machine learning techniques. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3015, 205–214. https://doi.org/10.1007/978-3-540-24668-8_21
- Memon, M. H., Li, J. P., Memon, I., & Arain, Q. A. (2017). GEO matching regions: multiple regions of interests using content based image retrieval based on relative locations. *Multimedia Tools and Applications*, 76(14), 15377–15411. <https://doi.org/10.1007/s11042-016-3834-z>
- Mensink, T., Verbeek, J., & Csurka, G. (2013). Tree-structured CRF models for interactive image labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2), 476–489.

Bibliography

- <https://doi.org/10.1109/TPAMI.2012.100>
- Mensink, T., Verbeek, J., Perronnin, F., & Csurka, G. (2012). Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7573 LNCS(PART 2), 488–501. https://doi.org/10.1007/978-3-642-33709-3_35
- Mezaris, V., Kompatsiaris, I., Strintzis, M. G., & Rd, K. T. (2003). *AN ONTOLOGY APPROACH TO OBJECT-BASED IMAGE RETRIEVAL* *Information Processing Laboratory Electrical and Computer Engineering Department Aristotle University of Thessaloniki Informatics and Telematics Institute Thessaloniki 57001, Greece.* 3–6.
- Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8828(c), 1–20. <https://doi.org/10.1109/TPAMI.2021.3059968>
- Monay, F., & Gatica-Perez, D. (2003). On image auto-annotation with latent space models. *Proceedings of the ACM International Multimedia Conference and Exhibition*, 275–278. <https://doi.org/10.1145/957052.957070>
- Monay, F., & Gatica-Perez, D. (2004). PLSA-based image auto-annotation: Constraining the latent space. *ACM Multimedia 2004 - Proceedings of the 12th ACM International Conference on Multimedia*, 348–351.
- Moran, S., & Lavrenko, V. (2014a). *A sparse kernel relevance model for automatic image annotation.* <https://doi.org/10.1007/s13735-014-0063-y>
- Moran, S., & Lavrenko, V. (2014b). Sparse kernel learning for image annotation. *ICMR 2014 - Proceedings of the ACM International Conference on Multimedia Retrieval 2014*, 113–120. <https://doi.org/10.1145/2578726.2578734>
- Moutarde, F. (2019). *IA: vers des robots intelligents ? October 2018.*
- Mueen, A., Zainuddin, R., & Baba, M. S. (2008). Automatic multilevel medical image annotation and retrieval. *Journal of Digital Imaging*, 21(3), 290–295. <https://doi.org/10.1007/s10278-007-9070-3>
- Murthy, V. N., Maji, S., & Manmatha, R. (2015). *Automatic Image Annotation using Deep Learning Representations.* 603–606.
- Nayak, S. R., Padhy, R., & Mishra, J. (2017). Texture analysis methods: A review. *Journal of Advanced Research in Dynamical and Control Systems*, 9(11), 46–52.
- Ning, Z., Zhou, G., Chen, Z., & Li, Q. (2018). Integration of Image Feature and Word Relevance: Toward Automatic Image Annotation in Cyber-Physical-Social Systems. *IEEE*

Bibliography

- Access*, 6, 44190–44198. <https://doi.org/10.1109/ACCESS.2018.2864332>
- Niu, Y., Lu, Z., Wen, J., Xiang, T., & Chang, S. (2019). Multi-Modal Multi-Scale Deep Learning for Large-Scale Image Annotation. *IEEE Transactions on Image Processing*, 28(4), 1720–1731. <https://doi.org/10.1109/TIP.2018.2881928>
- Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1717–1724. <https://doi.org/10.1109/CVPR.2014.222>
- Ouni, A., Royer, E., Chevaldonné, M., & Dhome, M. (2021). Leveraging semantic segmentation for hybrid image retrieval methods. *Neural Computing and Applications*, 3. <https://doi.org/10.1007/s00521-021-06087-3>
- Pagare, R., & Shinde, A. (2012). A Study on Image Annotation Techniques. *International Journal of Computer Applications*, 37(6), 42–45. <https://doi.org/10.5120/4616-6295>
- Pan, J. Y., Yang, H. J., Faloutsos, C., & Duygulu, P. (2004). GCap: Graph-based automatic image captioning. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2004-Janua*(January). <https://doi.org/10.1109/CVPR.2004.353>
- Park, S. B., Lee, J. W., & Kim, S. K. (2004). Content-based image classification using a neural network. *Pattern Recognition Letters*, 25(3), 287–300. <https://doi.org/10.1016/j.patrec.2003.10.015>
- Pass, G., & Zabih, R. (1996). Histogram refinement for content-based image retrieval. *IEEE Workshop on Applications of Computer Vision - Proceedings*, 96–102. <https://doi.org/10.1109/acv.1996.572008>
- Pedersen, K. S. (n.d.). *Salient Point and Scale Detection by Minimum Likelihood*. 59–72.
- Perronnin, F., & Dance, C. (2007). Fisher Kernels on Visual Vocabularies for Image Categorization. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. <https://doi.org/10.1109/CVPR.2007.383266>
- Putthividhy, D., Attias, H. T., & Nagarajan, S. S. (2010). Topic regression multi-modal Latent Dirichlet Allocation for image annotation. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3408–3415. <https://doi.org/10.1109/CVPR.2010.5540000>
- Qi, G., Liu, W., Aggarwal, C., & Huang, T. (2017). Joint Intermodal and Intramodal Label Transfers for Extremely Rare or Unseen Classes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7), 1360–1373. <https://doi.org/10.1109/TPAMI.2016.2587643>
- Qi, X., & Han, Y. (2007). Incorporating multiple SVMs for automatic image annotation. *Pattern*

Bibliography

- Recognition*, 40(2), 728–741. <https://doi.org/10.1016/j.patcog.2006.04.042>
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106. <https://doi.org/10.1007/bf00116251>
- Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys*, 28(1), 71–72. <https://doi.org/10.1145/234313.234346>
- R, J. V. C. I., Jin, C., & Jin, S. (2016). Image distance metric learning based on neighborhood sets for automatic image annotation. *JOURNAL OF VISUAL COMMUNICATION AND IMAGE REPRESENTATION*, 34, 167–175. <https://doi.org/10.1016/j.jvcir.2015.10.017>
- R, J. V. C. I., Rad, R., & Jamzad, M. (2017). Image annotation using multi-view non-negative matrix factorization with different number of basis vectors. *Journal of Visual Communication and Image Representation*, 46, 1–12. <https://doi.org/10.1016/j.jvcir.2017.03.005>
- R, J. V. C. I., Yang, Y., Zhang, W., & Xie, Y. (2015). Image automatic annotation via multi-view deep representation. *Journal of Visual Communication and Image Representation*, 33, 368–377. <https://doi.org/10.1016/j.jvcir.2015.10.006>
- Rad, R., & Jamzad, M. (2015). *Automatic image annotation by a loosely joint non-negative matrix factorisation*. 9, 806–813. <https://doi.org/10.1049/iet-cvi.2014.0413>
- Rosten, E., & Drummond, T. (2006). *Machine Learning for High-Speed Corner Detection*. 430–443.
- Rui, X. (2007). *Bipartite Graph Reinforcement Model for Web Image Annotation*. 585–594.
- S., M. B., P., S., & T., S. (2002). *Introduction to MPEG-7: multimedia content description interface*. [https://books.google.dz/books?hl=fr&lr=&id=CmSPGXF1yB4C&oi=fnd&pg=PR17&dq=Introduction+to+MPEG-7:+Multi-media+Content+Description+Language&ots=p5R6gJ3ZSa&sig=iMt8-mytfCoS6fVkrOE6JVxkRPY&redir_esc=y#v=onepage&q=Introduction to MPEG-7%3A Multi- media Conte](https://books.google.dz/books?hl=fr&lr=&id=CmSPGXF1yB4C&oi=fnd&pg=PR17&dq=Introduction+to+MPEG-7:+Multi-media+Content+Description+Language&ots=p5R6gJ3ZSa&sig=iMt8-mytfCoS6fVkrOE6JVxkRPY&redir_esc=y#v=onepage&q=Introduction+to+MPEG-7%3A+Multi-media+Conte)
- Sebe, N., Tian, Q., Louprias, E., Lew, M. S., & Huang, T. S. (2003). *Evaluation of salient point techniques*. 1–9. <https://doi.org/10.1016/j.imavis.2003.08.012>
- Semantic, O., For, A., & Indexing, M. I. (2007). *Modeling semantic aspects for cross-media image indexing*.
- Shi, J., & Malik, J. (1997). Normalized cuts and image segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 731–737. <https://doi.org/10.1109/cvpr.1997.609407>
- Shi, Z., Yang, Y., Hospedales, T. M., & Xiang, T. (2017). Weakly-Supervised Image Annotation and Segmentation with Objects and Attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2525–2538. <https://doi.org/10.1109/TPAMI.2016.2645157>

Bibliography

- Slabaugh, G., Unal, G., Wels, M., Fang, T., & Rao, B. (2009). Statistical Region-Based Segmentation of Ultrasound Images. *Ultrasound in Medicine and Biology*, 35(5), 781–795. <https://doi.org/10.1016/j.ultrasmedbio.2008.10.014>
- Song, H., Wang, P., Yun, J., Li, W., Xue, B., & Wu, G. (2020). A Weighted Topic Model Learned from Local Semantic Space for Automatic Image Annotation. *IEEE Access*, 8, 76411–76422. <https://doi.org/10.1109/ACCESS.2020.2989200>
- Song, L., Luo, M., Liu, J., Zhang, L., Qian, B., Li, M. H., & Zheng, Q. (2016). Sparse Multi-Modal Topical Coding for Image Annotation. *Neurocomputing*, 214, 162–174. <https://doi.org/10.1016/j.neucom.2016.06.005>
- Stangl, A., Morris, M. R., & Gurari, D. (2020). “Person, Shoes, Tree. Is the Person Naked?” What People with Vision Impairments Want in Image Descriptions. *Conference on Human Factors in Computing Systems - Proceedings*, 1–13. <https://doi.org/10.1145/3313831.3376404>
- Starck, J., Candès, E. J., & Donoho, D. L. (2000). *The Curvelet Transform for Image Denoising*. 1–27.
- Su, F., & Xue, L. (2015). Graph Learning on K Nearest Neighbours for Automatic Image Annotation. *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 403–410. <https://doi.org/10.1145/2671188.2749383>
- Suh, B., & Bederson, B. (2004). Semi-automatic image annotation using event and torso identification. ... of Maryland, College Park, Maryland, USA. <http://hcil.cs.umd.edu/trs/2004-15/2004-15.pdf>
- Talib, A., Mahmuddin, M., Husni, H., & George, L. E. (2013). A weighted dominant color descriptor for content-based image retrieval. *Journal of Visual Communication and Image Representation*, 24(3), 345–360. <https://doi.org/10.1016/j.jvcir.2013.01.007>
- Tamura, H., Mori, S., & Yamawaki, T. (1978). Textural Features Corresponding to Visual Perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6), 460–473. <https://doi.org/10.1109/TSMC.1978.4309999>
- Tang, J., Hong, R., Qi, G., Technologies, F., & Jain, R. (2011). *Image annotation by kNN-sparse graph-based label propagation over noisily tagged web images. August 2014*. <https://doi.org/10.1145/1899412.1899418>
- Tian, D., & Shi, Z. (2020). A two-stage hybrid probabilistic topic model for refining image annotation. *International Journal of Machine Learning and Cybernetics*, 11(2), 417–431. <https://doi.org/10.1007/s13042-019-00983-w>
- Tian, F., & Shen, X. (2015). Learning Label Set Relevance for Search Based Image Annotation. *Proceedings - 2014 International Conference on Virtual Reality and Visualization, ICVRV 2014*, 260–

Bibliography

265. <https://doi.org/10.1109/ICVRV.2014.39>
- Tommasi, T., Orabona, F., & Caputo, B. (2008). Discriminative cue integration for medical image annotation. *Pattern Recognition Letters*, 29(15), 1996–2002.
<https://doi.org/10.1016/j.patrec.2008.03.009>
- Tremeau, A., & Borel, N. (1997). A region growing and merging algorithm to color segmentation. *Pattern Recognition*, 30(7), 1191–1203. [https://doi.org/10.1016/S0031-3203\(96\)00147-1](https://doi.org/10.1016/S0031-3203(96)00147-1)
- Tuceryan, M., & Jain, A. K. (1993). texture analysis. In *Handbook of Pattern Recognition and Computer Vision* (pp. 235–276). https://doi.org/10.1142/9789814343138_0010
- Vailaya, A., Member, A., Figueiredo, M. A. T., & Jain, A. K. (2001). *Image Classification for Content-Based Indexing*. 10(1), 117–130.
- Vatani, A., Ahvanooy, M. T., & Rahimi, M. (2020). An effective automatic image annotation model via attention model and data equilibrium. *ArXiv*, 9(3), 269–277.
- Verma, Y., & Jawahar, C. V. (2012). Image annotation using metric learning in semantic neighbourhoods. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7574 LNCS(PART 3), 836–849.
https://doi.org/10.1007/978-3-642-33712-3_60
- Verma, Y., & Jawahar, C. V. (2017). Image Annotation by Propagating Labels from Semantic Neighbourhoods. *International Journal of Computer Vision*, 121(1), 126–148.
<https://doi.org/10.1007/s11263-016-0927-0>
- Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2017). *Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge*. 39(4), 652–663.
- Vizza, F., & Romani, D. (1809). Support vector machines for histogram-based image classification. *Ieee Transactions on Neural Networks*, 10(5), 1–9.
- Wang, X.L.; Hongwei, G.E.; Liang, S. (2018). Image automatic annotation algorithm based on canonical correlation analytical subspace and k-nearest neighbor. *Ludong Univ.*
- Wang, C. (2006). *Image Annotation Refinement using Random Walk with Restarts* *. 2–5.
- Wang, C., Jing, F., Zhang, L., & Zhang, H. J. (2006). Scalable search-based image annotation of personal images. *Proceedings of the ACM International Multimedia Conference and Exhibition*, 269–278. <https://doi.org/10.1145/1178677.1178714>
- Wang, C., Jing, F., Zhang, L., & Zhang, H. J. (2007). Content-based image annotation refinement. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1.
<https://doi.org/10.1109/CVPR.2007.383221>
- Wang, C., Yan, S., Zhang, L., & Zhang, H. J. (2009). Multi-label sparse coding for automatic

Bibliography

- image annotation. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, 2009 IEEE*, 1643–1650.
<https://doi.org/10.1109/CVPRW.2009.5206866>
- Wang, C., Zhang, L., & Zhang, H. J. (2008). Learning to reduce the semantic gap in web image retrieval and annotation. *ACM SIGIR 2008 - 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Proceedings*, 355–362.
<https://doi.org/10.1145/1390334.1390396>
- Wang, Hua, Huang, H., & Ding, C. (2011). Image annotation using bi-relational graph of images and semantic labels. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 793–800. <https://doi.org/10.1109/CVPR.2011.5995379>
- Wang, Huan, Jiang, X., Chia, L., & Tan, A. (2008). Ontology enhanced web image retrieval. *Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval - MIR '08*, 195. <https://doi.org/10.1145/1460096.1460128>
- Wang, J. Z., Li, J., & Wiederhold, G. (2001). *SIMPLcity: Semantics-Sensitive Integrated Matching for Picture Libraries*. 23(9), 947–963.
- Wang, R., Xie, Y., Yang, J., Xue, L., Hu, M., & Zhang, Q. (2017). Large scale automatic image annotation based on convolutional neural network. *Journal of Visual Communication and Image Representation*. <https://doi.org/10.1016/j.jvcir.2017.07.004>
- Wang, W., Hu, Y., Zou, T., Liu, H., Wang, J., & Wang, X. (2020). A New Image Classification Approach via Improved MobileNet Models with Local Receptive Field Expansion in Shallow Layers. *Computational Intelligence and Neuroscience*, 2020.
<https://doi.org/10.1155/2020/8817849>
- Wang, Y., Mei, T., Gong, S., & Hua, X. S. (2009). Combining global, regional and contextual features for automatic image annotation. *Pattern Recognition*, 42(2), 259–266.
<https://doi.org/10.1016/j.patcog.2008.05.010>
- Wei, W., Wu, Q., Chen, D., Zhang, Y., Liu, W., Duan, G., & Luo, X. (2021). Automatic image annotation based on an improved nearest neighbor technique with tag semantic extension model. *Procedia Computer Science*, 183(2018), 616–623.
<https://doi.org/10.1016/j.procs.2021.02.105>
- Wenyin, L., Dumais, S., Sun, Y., Zhang, H., Czerwinski, M., Field, B., & Way, O. M. (1999). *Semi-Automatic Image Annotation*.
- Weston, J., Bengio, S., & Usunier, N. (2010). *WSABIE: Scaling Up To Large Vocabulary Image Annotation*. 2764–2770.

Bibliography

- Wu, B., Chen, W., Sun, P., Liu, W., Ghanem, B., & Lyu, S. (2018). Tagging like Humans : Diverse and Distinct Image Annotation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7967–7975.
- Wu, L., Jin, R., & Jain, A. K. (2013). Tag completion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3), 716–727.
<https://doi.org/10.1109/TPAMI.2012.124>
- Xia, S., Chen, P., Zhang, J., Li, X., & Wang, B. (2016). *Utilization of rotation-invariant uniform LBP histogram distribution and statistics of connected regions in automatic image annotation based on multi-label learning*.
- Xu, X., Shimada, A., & Taniguchi, R. I. (2013). Image annotation by learning label-specific distance metrics. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8156 LNCS(PART 1), 101–110.
https://doi.org/10.1007/978-3-642-41181-6_11
- Xue, Z., Li, G., & Huang, Q. (2018). Joint multi-view representation and image annotation via optimal predictive subspace learning. *Information Sciences*, 451–452, 180–194.
<https://doi.org/10.1016/j.ins.2018.03.051>
- Y.I. Chang, B.Y. Yang, W. H. Y. (2003). A bit-pattern-based matrix strategy for efficient iconic indexing of symbolic pictures. *Pattern Recognition Letters*, 24(1–3), 537–545.
<https://doi.org/10.1016/j.patcog.2011.05.013>
- Yasuhide MORI, TAKAHASHI, H., & OKA, R. (1999). *Image-to-word transformation based on dividing and vector quantizing images with words*.
- Yavlinsky, A. (2007). *Department of Computing Image indexing and retrieval using automated annotation*. June.
- Yavlinsky, A., Schofield, E., & Rüger, S. (2005). Automated image annotation using global features and robust nonparametric density estimation. *Lecture Notes in Computer Science*, 3568, 507–517. https://doi.org/10.1007/11526346_54
- Yining Deng, Manjunath, B. S., & Shin, H. (1999). Color image segmentation. *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, Vol:2, For, 446–451. <https://doi.org/10.1109/CVPR.1999.784719>
- Yogamangalam, R., & Karthikeyan, B. (2013). Segmentation techniques comparison in image processing. *International Journal of Engineering and Technology*, 5(1), 307–313.
- Zamiri, M., & Sadoghi Yazdi, H. (2021). Image annotation based on multi-view robust spectral clustering. *Journal of Visual Communication and Image Representation*, 74(December 2020),

Bibliography

103003. <https://doi.org/10.1016/j.jvcir.2020.103003>
- Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8689 LNCS(PART 1), 818–833. https://doi.org/10.1007/978-3-319-10590-1_53
- Zhang, D., Islam, M. M., & Lu, G. (2012). A review on automatic image annotation techniques. *Pattern Recognition*, 45(1), 346–362. <https://doi.org/10.1016/j.patcog.2011.05.013>
- Zhang, D., Islam, M. M., Lu, G., & Hou, J. (2009). Semantic Image Retrieval Using Region Based Inverted File. *2009 Digital Image Computing: Techniques and Applications*, 242–249. <https://doi.org/10.1109/DICTA.2009.48>
- Zhang, D., & Lu, G. (2004). *Review of shape representation and description techniques*. 37, 1–19. <https://doi.org/10.1016/j.patcog.2003.07.008>
- Zhang, D., Wong, A., Indrawan, M., & Lu, G. (n.d.). *Content-based Image Retrieval Using Gabor Texture Features*.
- Zhang, J., Gao, Y., Feng, S., Yuan, Y., & Lee, C. H. (2016). Automatic image region annotation through segmentation based visual semantic analysis and discriminative classification. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2016-May*, 1956–1960. <https://doi.org/10.1109/ICASSP.2016.7472018>
- Zhang, J., Mu, Y., Feng, S., Li, K., Yuan, Y., & Lee, C. H. (2018). Image region annotation based on segmentation and semantic correlation analysis. *IET Image Processing*, 12(8), 1331–1337. <https://doi.org/10.1049/iet-ipr.2017.0917>
- Zhang, J., Tao, T., Mu, Y., Sun, H., Li, D., & Wang, Z. (2019). Web image annotation based on Tri-relational Graph and semantic context analysis. *Engineering Applications of Artificial Intelligence*, 81(June 2018), 313–322. <https://doi.org/10.1016/j.engappai.2019.02.018>
- Zhang, J., Zhao, Y., Li, D., Chen, Z., & Yuan, Y. (2015). A novel image annotation model based on content representation with multi-layer segmentation. *Neural Computing and Applications*, 26(6), 1407–1422. <https://doi.org/10.1007/s00521-014-1815-6>
- Zhang, Rui, Guan, L., Zhang, L., & Wang, X. J. (2011). Multi-feature pLSA for combining visual features in image annotation. *MM'11 - Proceedings of the 2011 ACM Multimedia Conference and Co-located Workshops*, 1513–1516. <https://doi.org/10.1145/2072298.2072053>
- Zhang, Ruofei, Zhang, Z., Li, M., Ma, W. Y., & Zhang, H. J. (2006). A probabilistic semantic model for image annotation and multi-modal image retrieval. *Multimedia Systems*, 12(1), 27–33. <https://doi.org/10.1007/s00530-006-0025-1>

Bibliography

- Zhang, W., Hu, H., & Hu, H. (2018). Training Visual-Semantic Embedding Network for Boosting Automatic Image Annotation. *Neural Processing Letters*.
<https://doi.org/10.1007/s11063-017-9753-9>
- Zhang, W., Hu, H., Hu, H., & Yu, J. (2020). Automatic image annotation via category labels. *Multimedia Tools and Applications*, 79(17–18), 11421–11435. <https://doi.org/10.1007/s11042-019-07929-y>
- Zhang, X., & Liu, C. (2015). Image annotation based on feature fusion and semantic similarity. *Neurocomputing*, 149, 1658–1671. <https://doi.org/10.1016/j.neucom.2014.08.027>
- Zhu, S. C., & Yuille, A. (1996). Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(9), 884–900.
- Zhu, Z., & Hangchi, Z. (2020). Image annotation method based on graph volume network. *Proceedings - 2020 International Conference on Intelligent Transportation, Big Data and Smart City, ICITBS 2020*, 885–888. <https://doi.org/10.1109/ICITBS49701.2020.00195>

Glossary

A.

AIA: Automatic Image Annotation,

ANNs : Artificial Neural Networks

B.

BG : Bi-relational Graph

BoVW : Bag of Visual Words

BRIEF : Binary robust independent elementary features

C.

CCV : color coherence vector

CCV :color coherence vector

CIAR : content-based image annotation refinement

CLD :color layout descriptor

CM : color moments

CMRM : Cross-media Relevance Models

CNN : Convolutional Neural Network

CR : Causal Relationships

CRF : conditional random fields

CRM : Continuous-space Relevance Model

CSD : color structure descriptor

D.

Glossary

D2IA : diverse and distinct image annotation

DCD : dominant color descriptor

DCMRM :dual cross-media relevance model

DCT : Discrete Cosine Transform

DNN : Deep Neural Networks

DPP : determinantal point process

F.

FAST : features from accelerated segment test

G.

GAN : generative adversarial network

H.

HMM : Hidden Markov Model

J.

JEC : Joint Equal Contribution

JSEG : image segmentation

N.

N2D clustering : Not too deep clustering

K.

KCCA : Kernel Canonical Correlation Analysis

KML : kernel metric learning

L.

LDA : Latent Dirichlet Allocation

LL-PLSA : local learning-based PLSA

M.

MBRM : Multiple-Bernoulli Relevance Model

MKL : Multiple kernel refinement

ML : machine learning

mm-pLSA : multilayer multimodal probabilistic Latent Semantic Analysis

MV : Multitask Voting

N.

NCM : nearest class mean

NCut : normalised cut

NN : Nearest-neighbor

NS : normalized score

NSC : Nearest Span Series

P.

PLSA: Probabilistic Latent Semantic Analysis model

PLSA-WORDS : Probabilistic Latent Semantic Analysis model WORDS

PTM: probability topical model

R.

RKML : robust kernel metric learning

RNN : Recurrent Neural Network

RWR :Random Walk with Restart

S.

SAMSI : Semantic annotated Markovian Semantic Indexing

Glossary

SIFT : Scale-invariant feature transform

SCD : scalable color descriptor.

SKL-CRM : Sparse kernel relevance model

SL : Supervised learning

SMMTC: sparse multi-modal topical coding

STC : sparse topical coding.

SURF : speeded-up robust features

SVM : support vector machine

T.

tr-mmLDA : topical regression multi-modal Latent Dirichlet Allocation

V.

VA-Files :Vector Approximation Files

VLAD : Vector of Locally Aggregated Descriptors

2.

2PKNN : two-pass k-nearest neighbor

