POPULAR DEMOCRATIC REPUBLIC OF ALGERIA
HIGHER EDUCATION AND SCIENTIFIC RESEARCH'S MINISTRY

Faculty of Exact Sciences, Nature Sciences and Life
Department of Mathematics and Computer Science
Larbi Ben M'hidi University, Oum El Bouaghi, Algeria

Semantic extraction and interpretation of image content

Submitted submitdate, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science

ZGA Adel

Supervised by:
Professor **Nini Brahim**

Jury members:

Dr. **Zertal Soumia**	**Larbi Ben M'hidi Univ. of Oum El-Bouaghi**	**President**
Dr. Menassel Rafik	**Larbi Tbessi Univ. of Tebessa**	**Examiner**
Dr. Lakhdar.laimeche	**Larbi Tbessi Univ. of Tebessa**	**Examiner**
Dr. Khaldi Amine	**Kasdi Merbah Univ. of Ouargla**	**Examiner**
Pr. NINI Brahim	**Larbi Ben M'hidi Univ. of Oum El-Bouaghi**	**Supervisor**

September 2021

Abstract

Nowadays, three challenges of relationship-detection should be considered in order to build a strong model namely; long-tail problem, large intra-class divergence, and the semantic dependency or semantic gap. The aim of content-based image retrieval systems must provide maximum support in bridging the semantic gap between the simplicity of available visual features and the richness of the user semantics. Another issue is the long-tail problem where there is a low infrequent appearance of some objects (i.e. predicates) versus to the high occurrence of others. For that, an adequate scaling is demanded. The third problem that must be solved to build a strong CBIR system is the Intra/inter-class divergence. For the Intra-class divergence, objects (i.e., predicates) are belonging to the same class but can't be represented with the same visual characteristics, whereas the inter-class divergence is where the similar visual descriptors can relate to two objects (i.e., predicates) that are not related to each other. In order to overcome those challenges, we propose three main contributions: 1) an ontological semantic model to filter false negatives/positives using a statistical ranking module. 2) the combination of semantic ontological module and visual relationship module that both takes as input the results of the statistical ranking module and produces as output classification of <human−predicate−object>. 3) a semantic model for the visual relationship module that ranks the prediction of relation classes by transferring the spatial relationship onto a high dimension spatial feature. Finally, we used HCVRD that highlights two important practical problems, the long-tail distribution issue, and the zero-shot problem. The experimental results on the HCVRD dataset demonstrate the superior performance of the proposed approach.

Keywords

Deep learning; Semantic Gap; Ontologies; Human-object interaction; large Intra/inter class divergence; Long tail problem; Content based image retrieval

ملخص

اليوم في التفسير الدلالي لمحتوى الصور توجد تحديات يجب ان تؤخذ بعين الاعتبار وذلك من اجل بناء نموذج قوي وتتمثل هذه التحديات في مشكلة الذيل الطويل و التباعد الكبير داخل الطبقة والتبعية الدلالية أو الفجوة الدلالية. الهدف من أنظمة البحث عن الصور القائمة على المحتوى هو توفير أقصى قدر من الدعم لسد الفجوة الدلالية بين بساطة الميزات المرئية المتاحة وثراء دلالات المستخدم، مشكلة أخرى هي مشكلة الذيل الطويل حيث يوجد تكرار منخفض لبعض الأشياء (أي المسندات) مقارنة بالظهور المرتفع للآخرين ولهذا التحجيم المناسب مطلوب. الملة الشكّالثة التي يجب حلها لبناء نظام $CBIR$ قوي هي الاختلاف داخل و بين الطبقات. بالنسبة للاختلاف داخل الطبقة ، تنتمي الكائنات (أي المسندات) إلى نفس الفئة ولكن لا يمكن تمثيلها بنفس الخصائص المرئية ، في حين أن الاختلاف بين الفئات هو النقطة التي يمكن أن ترتبط فيه الواصفات المرئية المماثلة بكائنين (أي المسندات). من أجل التغلب على هذه التحديات ، نقترح ثلاث مساهمات رئيسية: ١) نموذج دلالي وجودي لتصفية السلبيات / الإيجابيات الكاذبة باستخدام وحدة التصنيف الإحصائي. ٢) الجمع بين الوحدة الأنطولوجية الدلالية ووحدة العلاقات المرئية اللذان يأخذان كمدخلات نتائج وحدة التصنيف الإحصائي وينتجان كمخرجات تصنيف < الإنسان المسند الكائن >. ٣) نموذج دلالي لوحدة العلاقة المرئية يصنف التنبؤ بفئات العلاقة عن طريق نقل العلاقة المكانية إلى خاصية مكانية كبيرة الأبعاد. أخيرًا ، استخدمنا HCVRD الذي يسلط الضوء على مشكلتين عمليتين مهمتين ، مشكلة توزيع $long - tail$ ومشكلة $zero - shot$. اوضحت النتائج التجريبية على مجموعة بيانات HCVRD ان الطريقة المقترحة خققت درجة عالية من الدقة

كلمات رئيسية

تعلم عميق؛ الفجوة الدلالية علم الوجود تفاعل الكائن البشري تباعد كبير داخل / بين الفئات ؛ مشكلة طويلة الذيل استرجاع الصور على أساس المحتوى

Résumé

De nos jours, trois défis de détection de relation doivent être prises en considération afin de construire un modèle fiable qui sont ; le problème de long-tail, la large divergence intra-classe et la dépendence sémantique ou le trou sémantique. L'objectif des systèmes de recherche basée sur le contenu doit fournir un support maximal pour combler le trou sémantique entre la simplicité des caractéristiques visuelles disponibles et la richesse des sémantiques de l'utilisateur. Un autre problème qui est le long-tail problème où il y a une faible apparition peu fréquente de certains objets (c'est-à-dire les prédicats) versus à l'occurrence élevée des autres. Pour cela, une mise à l'échelle est exigée. Le troisième problème qui doit être résolu pour construire un système CBIR fiable est bien la divergence Intra/inter-classe. Pour la divergence Intra-classe, les objets (c'est-à-dire les prédicats) appartiennent à la même classe mais ils ne peuvent pas être représentés avec les mêmes caractéristiques visuelles, tandis que la divergence inter-classe est celle où les descripteurs visuels similaires peuvent se rapporter à deux objets (c'est-à-dire les prédicats) qui ne sont pas liés entre eux. Dans le but de surpasser ces défis, on propose trois principales contributions : 1) un modèle d'ontologie sémantique pour filtrer les faux négatifs/positifs en utilisant un module de classement statistique. 2) la combinaison de module d'ontologie sémantique et le module de relation visuelle qui prennent tous les deux comme entrée les résultats de module de classement statistique et produisent comme sortie la classification de <homme−prédicat−objet>. 3) un modèle sémantique pour le module de relation visuelle qui classifie la prédiction des classes de relation en transférant la relation spatiale sur une caractéristique spatiale de haute dimension. Finalement, nous avons utilisé HCVRD qui montre deux problèmes pratiques très importants, le problème de distribution de longue-tail, et le problème de zero-shot. Les résultats expérimentaux sur la base HCVRD montre la supériorité des performances de notre approche proposée.

Mots clés

Apprentissage profond ; fossé sémantique ; les ontologies ; interaction homme-objet ; large divergence Intra/inter classe, le problème de long-tail ; recherche d'image basée sur le contenu.

Acknowledgements

Any attempt at any level can 't be satisfactorily completed without the support and guidance of **ALLAH**, may Allah accept this work also.

I might want to express my unique thanks of appreciation to my educator "**Pr. NINI Brahim**" who offered me the brilliant chance to do this superb Doctoral thesis, which likewise assisted me with lot of Exploration and I came to know about such countless new things I'm truly appreciative to them. Without you, this thesis wouldn't be realised.

For Every one who reads this thesis.

Thank you,

Adel Zga

Contents

List of Figures

List of Tables

Glossary

ACC Accuracy

AI Artificial Intelligence

ANNs Artificial Neural Networks

BOW Bag of Words

CBIR Content Based Image Retrieval

CNN Convolutional Neural Network

CNNs Convolutional Neural Networks

CSS Curvative Scale Space Descriptors

DC Domain Consensus

DF Document Frequency

DIKW Data-Information-Knowledge-Wisdom

DL Deep Learning

DL Web Ontology Language

DNN Deep Neural Network

FC Fully Connected

FE Frame Element

FN False Negatives

FP False Positives

FPR False Positive Rate

FWIoU Frequency Weighted Intersection over Union

GMM Gaussian Mixing Model

GUI Graphical User Interface

HCVRD Large-scale human-Centric Visual Relationship Fetection dataset

HVS Hue, Saturation, Lightnes

ICA Independent Component Analysis

ILSVRC ImageNet Large Scale Visual Recognition Challenge

IoU Intersection Over Union

IoU Intersection of Union

IR Information Retrieval

ISO International Organization for Standardization

K-NN K-Nearest Neighbors

KBS Knowledge Based Systems

KI Knowledge Engineering

KNN k-Nearest Neighbors Algorithm

LBP Local Binary Pattern

LDA Linear Discriminant Analysis

LSI Latent Semantic Indexing

LU Lexical Units

MA Mean Accuracy

MAE Mean Absolute Error Loss

MDA Model Driven Architecture

MDD Model Driven Develop

MIoU Mean Intersection over Union

MSE Mean Squared Error Loss

MSE Mean Squared Error

MSRA MicroSoft Research Asia

NMS non-maximum suppression

NS Normalized Score

OLED OntoUML lightweight Editor

OWL Description Logics

PA Pixel Accuracy

PC principal Components

PCA Principal Component Analysis

PPV Positive Prediction Value

RBF Radial Basis Function

ReLU REctified Linear Unit

Resnet50 Pretrained model Resnet50

RGB Red Green Blue

RNN Recurrent Neural Network

RNN Recurrent Neural Networks

SE Software Engineering

SEI Semantic Extraction and Interpretation

SMC Simple Matching Coefficient

SMI Stanford Medical Informatics

SRD Semantic Relationship Detection

SVD singular Value Decomposition

SVM Support Vector Machine

SVM Support Vector Machines

TC Term Contribution

TF Term Frequency

TF-IDF Term Frequency Inversed Document Frequency

TF-RIDF Term Frequency Residual Inverse Document Frequency

TN True Negatives

TP True Positives

TVQ Term Variance Quality

URI Universal Resource Identifiers

VGG Pretrained model VGG

XML EXtensible Markup Language

1

General Introduction

1.1 Backgrounds and issues in CBIR

With the mass explosion of multimedia data, the development of search engines' applications to exploit it, becomes crucial. Within the framework of the thesis, increasingly applications that rely on computer vision must be endowed with several functions able to provide a semantic extraction and interpretation (SEI) of media content. In SEI content-based images, objects, and inter-object relationships (i.e., predicates) are first labeled with specific terms to form the general context of the image in question. The goal of indexing systems is to allow a user to find, in databases, all images which are much similar to an image in question (i.e., video, audio, text… etc.). An indexing program is conceived as a system that takes as an entry a reference image and which gave as a result a set of images similar to this one. This allows them to be sorted from most similar to least similar according to the labels defined by the context of the image in question.

Historically, the first type of method proposed relies on "searches by example": where an input image will be compared to all images in the database, a list of images is then sorted in terms of resemblance to the request. More precisely, it is not the image's pixels that are compared, but "descriptors". These descriptors are often "low level", that is, they characterize aspects close to the raw signal. Comparisons between these descriptors are based on measures of distance. Over time, research by example has reached its limits, not only because of the modality in which the request is expressed, but it opens the possibility of expressing complex semantic requests, to which it turns out to be very difficult to answer via the existing researches. Moreover, the major difficulty linked to this case relies on the task of correspondence between the textual terms composing the query and the numerical values encoding the images; that is called "content interpretation". In computer vision, interpretation is defined as the process by which a scene, represented by one or more sources of information, is described automatically by its semantic content using a computer system.

However, the automatic annotation of objects and the inter-object relationship is a difficult task. Given N objects and R predicates, an inter-object relationship indexing system has to examine ($O(N^2 \times R)$) relations. These would lead to a huge number of potential relation types in real-world applications. For example, there exist more than 75K predicates in the Visual Genome dataset (1). However, there exist in literature three challenges in which labeling systems or indexing systems faced during interpreting the inter-object relationship. Overcoming those issues is key of building a strong model (2, 3, 4) that automatically interprets the content of each part of image in terms of context and inter-context relationship. The three problems are described in the following subsection within the framework of the thesis.

1.1.1 Semantic gap and dependency

The goal of content-based picture recovery frameworks should offer greatest help in connecting the semantic hole between the effortlessness of accessible visual elements and the lavishness of the client semantics (5). The semantic gap is what defines the apartments of the raw (tables of numbers) and semantic (context and inter-concept relationships) representations of an image (6, 7, 8). Based on (9), it is the absence of occurrence between the data that one can separate from the visual information and the understanding that similar information have for a client in a given circumstance. It is also accentuated because of large intra-class divergence and large inter-class divergence. An example of a semantic gap is given in Figure 1.1, the VGD system detected a <woman-on-pants>, semantically, the prediction should be <woman-wearing-pants>.

These problems make it harder for a machine to deduce whether the image content corresponds to the context labels or not. Reducing the semantic gap is one of the major difficulties of an automatic system of annotation/indexing of content-based images. The computer vision and auto-indexing communities continue to address this issue. Many types of research have been done with the aim of increasing the correlation between semantically similar visual content by providing good descriptors (10, 11). Other machine learning methods (12, 13) proposed a matching between low-level features and their effectively semantic descriptions or concepts. However, an improvement is achieved, but not at the level of semantic interpretation of humans.

1.1.2 Long-tail problem

The long-tail problem (14), is due to the unbalanced classes in the same dataset. As consequence, there will be a low infrequent appearance of some objects (i.e. predicates) versus the high occurrence of others. For that, adequate scaling is demanded. Indeed, there is quite a

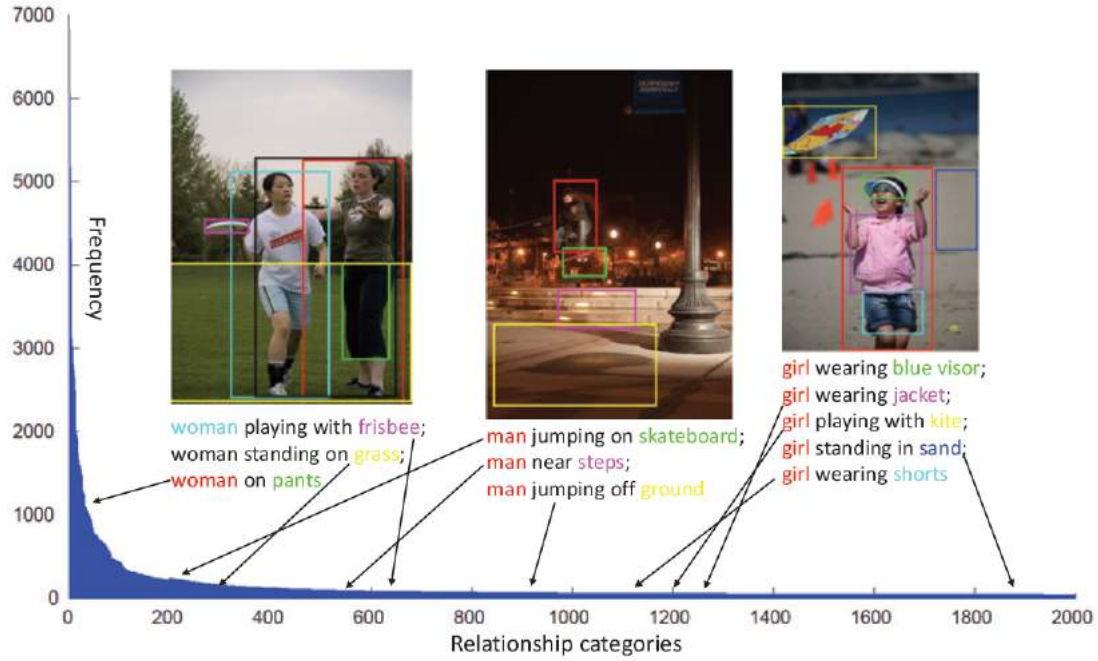


Figure 1.1: An example of long-tail problem in HCVRD dataset (4)

lot of learning methods that are annotated using only one or two classes. Table 1.1 represents the distribution of positive and negative samples as well as their ratio, for some concepts in the TRECVid 2012 collection (15). As it is noted, there are few concepts that have more positive samples than negative ones, and these kinds of concepts are often generic, and they, therefore, obtain their high frequency via the occurrence of the concepts which are specific to them, as for example the "Person" which is the head concept of several other bottom concepts: Male, Female, Child, Boy, Girl, Actor, etc. An occurrence of one of these bottom concepts necessarily implies the occurrence of their head concept. As a result, for majority classes, we will have a hundred positive examples vs. hundreds of thousands of negative examples. One can ask, why don't we scale samples of the minority classes to balance the samples of the majority classes? The answer is: a good model means an optimized accuracy or error rate; and to achieve this, the optimization is biased indeed, but by the minority class samples. On the other hand, when seeking for scaling between samples of the majority and minority classes, few examples of majority class are used to learn the model. For that, it will undoubtedly unable to generalize or to scale the examples of each class. The long-tail problem occurs due to the optimization process imposed by standard learning and classification algorithms and also, due to the insufficient samples of classes. In Figure 1.1, we illustrate an example of the long-tail problem, Only the top-2000 relationships are shown because the tail is too long. Three example images are also shown, where the shade of people and articles

in the expressions compare to the shade of the bouncing boxes. The arrows indicate the 'location' of the relationship in the label distribution. As we can see, most of the context (i.e., object and inter-object relationship) lie in the tail. Some of them such as 'girl wearing blue visor' is not even in the top-2000 (4).

1.1.3 Large intra/inter-class divergence

Intra/inter-class divergence are two other problems that have a huge impact on the semantic information extraction in images. For the Intra-class divergence, objects (i.e., predicates) belong to the same class but can't be represented with the same visual characteristics; in another term, it is the visual variations or multiple representations of the same object. An example is illustrated in the following figure (Figure 1.2), flying airplane is different from a broken airplane!

Figure 1.2: Difference between the visual characteristics of the same object (example of the airplane class)

The inter-class divergence is where similar visual descriptors can relate to two objects (i.e., predicates) that are not related to each other; we talk here about the similar visual descriptors for two different objects. An illustrative example is shown in Figure 1.3, one descriptor can't be represented to different classes: car or house!

Figure 1.3: Visual ambiguity. The same visual content or two similar visual content can refer to two different meanings

In automatic annotation systems, inter/intra-class divergence problem (16) can leads to

a huge semantic gap problem as well. Spatial relationships are in the shape of: on, under, in, etc., if predicates labeling was incorrect, the retrieved semantic information will be ambiguous and therefore, interpreted wrongly. One can solve semantic gap, and other, long-tail problem or inter/intra-class divergence, but the smartest is to evoked and evolved the three issues when trying to build a strong semantic information retrieval system.

1.2 Context and Problematic

Huge number of researches addressed the area of content-based information retrieval (CBIR) (17, 18, 19, 20). But the difficulties and challenges of semantic indexing presented above (see Session .1.1) have not yet been solved. The issues related to the semantic gap are more targeted compared to the long tail problem and the intra/inter-class divergence, but it still not completely overcame. However, using low-level descriptors through sophisticated algorithms cannot effectively model the semantics of images. Indeed, this approach has many limitations, especially when dealing with large data sets (21, 22) where there is no direct link between the visual appearance and the semantic interpretation, more precisely, between low-level features and high-level semantic concepts (23, 24).

Although another problem in CBIR is related to the performance of semantic annotation systems where they seem to be improved, this improvement does not concern all concepts. Figure 1.1 shows the performance obtained for some concepts by the HCVRD relationship detection system. We can notice that there are concepts that we can detect effectively and for some others, performance is average, but current indexing systems are reported to have failed to detect certain contexts that give poor results. This is due to the indexing methods where concepts are treated independently of each other. This isn't ideal in light of the fact that a similar idea can show up in altogether different settings and can likewise be changing in a similar setting too.

Context can also be thought of as another concept or a set of other concepts and we will fall into the ambiguity of identifying a concept. One can solve this by having multiple descriptions of the detection concept, a rich background of each concept should be built. Another solution is the use of inter-concept relationship where it is supported by the fact that, an image is very rich in semantics, and that often a concept does not appear alone in an image, but the use of the low-level features to interpret the semantics won't fit the purpose. The inter-concept relationships semantic represents a very important source of information that should not be neglected. For that, an automatic system that relies on the visual appearance and the semantic between objects in images is needed.

It is important to emphasize that a human uses certain semantics and links between concepts in their reasoning to infer the occurrence of certain concepts in an image. Indeed,

the human does not need to look for the shape of an object to deduce its presence, he can use for that the occurrence of certain other semantics (e.g., The occurrence of a bed excludes the appearance of an airplane). Also, the human can build a strong deep, and rich background of concepts (i.e., object or predicate) in an image, with well-defined descriptions (e.g., a person can be described by age, sex, dress, activities ...). On the other hand, trying to extract and interpret the semantic of a concept on your own in an image is not the best approach, but a reinforcement of human vision in an automatic system could fit the purpose.

1.3 Objectives of the thesis

Before exploiting the semantic contextual information in the image in question, it is necessary to define what the context is. Indeed, the context does not have a precise definition and has been used in the state of the art in different fields in different ways. It is therefore essential to provide a definition that corresponds to our research problem.

This thesis is dedicated to the semantic extraction and interpretation of image content. The main goal is to provide approaches that exploit the semantic contextual information as well as the inter-concept relationship. This can't be done unless defining the context of each object belonging to the image in question. The search of the context of each object and their visual relationship with other objects belonging to the same image was the method adopted by many types of research done until now. But it is above all necessary to develop an automatic semantic interpreting system that relies on human semantic interpretation. In our work, we benefit from the advances in ontology. Ontologies are "an explicit specification of a conceptualization" (25). They ensure a mutual perspective of a specific space, just as a conventional model that is amiable to unaided machine handling (26).

Our thesis work has several strengths. The first important point is the genericity of our approaches. Indeed, our work is done in such a way as to achieve our objectives while proposing generic approaches, which may be addressed to any indexing system for the capturing any target concept. On the other hand, we insisted that our contributions should not be specific to a particular category of concepts. Another important point of this thesis work is the fact of considering several approaches acting at different levels of an indexing system. This would make it possible to compare and find the most suitable level to effectively exploit the context.

1.4 Contributions and thesis overview

In this thesis, we propose the use of approaches that define the context of each object based on the collection of objects belonging to the image in question. Instead of considering concept

samples independently, we propose the use of the collections of objects in the image. An object cannot be a part of a certain collection unless it was a domain of subsumption of the context subclass. We also propose the use of the cardinality constraints which are imposed between the inter-concept relationship (or., the relationship between objects). The strengthening of our approaches came from the benefits of using ontologies and deep learning to achieve not only the semantic contextual information of images, but to reduce the semantic gap, improve intra/inter-class divergence problem, and help in avoiding the problems that occur from the datasets that suffer from the long tail problem.

To summarize, the main contributions of this paper include:

1. We use the HCVRD dataset that highlights two interesting issues, the first is the long-tail distribution issue, and the second is the zero-shot problem. However, our work is not specific to the HCVRD dataset.

2. We use ontology learning layer cake to build an ontological model that describes and transforms the visual interpretation into a semantic interpretation of each object.

3. We propose the ontological semantic model to filter $false\ positive/negative$. We compare the object class proposals (i.e., probabilities of classification that are extracted from the object detection module) with information extracted from the ontological formal description using the statistical ranking module.

4. We propose the combination of semantic ontological module and visual relationship module that both take as input the results of the statistical ranking module and produce as output classification of <human-predicate-object>. The semantic ontological module gives a rich semantic rank of prediction between connecting objects. Also, the visual relationship module ranks the prediction of relation classes by transferring the spatial relationship onto a high dimension spatial feature.

The main body of this thesis is divided as follows.

- Chapter 2 describes computer vision and information retrieval in images. This chapter contains two parts. The first part is dedicated to present image representation, descriptors, and segmentation techniques, annotation techniques, search techniques, and techniques recommendation. The second part defines machine learning, classification, and optimization.

- Chapter 3 is a state of the art about ontology learning model. This chapter is also divided into two main parts, the first part is dedicated to giving an overview of the ontology learning models whereas, in the second part, we will detail the used learning strategy which is the ontology learning "layer cake".

- Chapter 4 describes the visual relationship extraction in images and a semantic interpretation ranking with ontologies. In this chapter a detailed explanation about the proposed methods. We start by giving motivations and contributions to the selected research field. After that, a demonstration and experimentation under some assumptions are done in order to illustrate the results obtained.
- Chapter 5 summarizes this thesis

2

Computer Vision and Information Retrieval in images

2.1 Introduction

Content-Based Image Retrieval (CBIR) is to find images based on their visual characteristics. Before you can search for images, you must first extract their characteristics. Extracting the features contained in an image is called structural description. This can take the form of an image or any data structure allowing a description of the entities contained in the image. Essentially, image analysis involves segmentation where we will try to associate each region of the image with a label based on the information carried and the spatial distribution. Images are conventionally described as reflecting their low-level characteristics such as texture, color, shape, etc. A typical case of utilizing low-level features is looking for images that are outwardly like a given inquiry model (image).

This technique is opposed to keyword search for images, which was historically offered by search engines where images are found using accompanying text rather than the content of the image itself [1]. This chapter contains two parts. The first part is dedicated to present image representation, descriptors, and segmentation techniques, annotation techniques, search techniques, and techniques recommendation. The second part defines machine learning, classification, and optimization.

2.2 Computer Vision and Image Analysis

Computer vision is an interdisciplinary logical field that arrangements with how Computers can acquire undeniable level comprehension from computerized pictures or recordings. According to the point of view of designing, it looks to comprehend and robotize undertakings

[1] https://en.wikipedia.org/wiki/Content-based_image_retrieval

that the human visual framework can do (27, 28, 29). In this section, we will give more details about the computer vision as well as image representation.

2.2.1 Computer Vision

Computer vision errands incorporate techniques for procuring, handling, investigating, and understanding computerized pictures and extraction of high-dimensional information from this present reality to deliver mathematical or emblematic data, for example in the types of choices (30, 31, 32, 33). Understanding in this setting implies the change of visual pictures (the contribution of the retina) into portrayals of the world that check out to perspectives and can inspire fitting activity. This picture comprehension can be viewed as the unraveling of representative data from picture information utilizing models developed with the guide of math, material science, insights, and learning hypothesis (34).

2.2.2 Image Representation

2.2.2.1 Physical Representation

Traditional approaches for image searching are to represent them with low-level characteristics (35). Physical characteristics are the basis of image content search systems. The features that extracted from image represent global features, and features obtained from a region of the image are called local features. This latter (local) characteristics were found to be closer to perception and are based on the images segmentation into regions.

- **The color**

Colors are one of the most broadly utilized characteristics in content-based image retrieval systems. They are rich in information and widely used for the representation of the image. They form a significant part of human vision. Human perception of color in an image is a complex and subjective process (36). Indeed, this data varies considerably with the orientation of the surfaces, the camera, and the illumination (positions and wavelength of light sources), for example. It is possible to represent the color in different spaces namely RGB (Red Green Blue), HVS (Hue, Saturation, Lightness), etc. The most common is undoubtedly the RGB space which codes the color of a pixel on a three-dimensional vector.

- **The texture**

The texture is the second widely used visual characteristic in CBIR. It makes it possible to fill a void that color is incapable of filling, especially when the color distributions are very close. The texture is generally defined as the repetition of a pattern creating a visually

homogeneous image. It can be seen as a set of spatially arranged pixels, thus creating a homogeneous region. Researches that use low-level characteristics, texture plays a very important role because it gives significant data in the characterization of pictures since it portrays the substance of many pictures such as the skin of organic products, trees, blocks, and texture, and so forth Hence, the surface is a significant component in the meaning of undeniable level semantics.

- **The form**

The shape makes it possible to extract particularly robust and discriminating characteristics. It is often called characteristics of objects because they presuppose image segmentation and therefore manipulate information related to real objects. Like texture, the shape is a complementary characteristic of color. CBIR systems have focused on extracting geometric attributes. Two techniques have been proposed to represent shapes: descriptors based on regions and descriptors based on contours (boundaries). Region-based descriptors are used to characterize the entire shape of a region, and border-based descriptors focus on the outlines of shapes (36). Shapes are very useful in certain areas such as in man-made objects. Shapes are difficult to apply to color images used in most documents.

- **Spatial localization**

In addition to the aforementioned characteristics, it is also possible to consider the spatial organization of the various primitives as a description as such. It is obvious that such a description constitutes an intermediate level between the raw low-level characteristics and the interpretation of the images, and that this level can be very expressive. Thus, when a user searches for an image that represents a complex object,

In summary, we can say that the low-level features of the image can be either extracted from the entire image or from parts of the image. Most current systems focus on image regions because searches based on overall image characteristics are relatively straightforward, but region-based searches are efficient because they are closer to human perception.[1]

2.3 Image Descriptors

2.3.1 Local Descriptors and Global Descriptors

One can use descriptors characterizing the entire image (global descriptor) or several local descriptors each characterizing a part of the image. Modern techniques in imaging tend to privilege local descriptors over global ones because local descriptors are more efficient and

[1] http://glotin.univ-tln.fr/MCBIR/Segmentation_images_principes.pdf

they allow a finer search and better absorb certain variations. In the case of global descriptors, a single descriptor writes the entire image, this will be robust against noise.

The disadvantage of these descriptors is that they do not make it possible to distinguish parts of the images, for example; the objects in the image, (except in the case where the image contains only one single object in a plain background). In contrast, the local descriptors are associated with a part/region of the image that we first detect before calculating the descriptor, this part can relate to an object.

2.3.2 Visual Descriptors

Descriptors can also be categorized according to the type of modality they represent: visual descriptors, audio descriptors, motion descriptors, etc. These points are detailed in the following subsections.

2.3.2.1 Color Descriptors

Color descriptors are widely utilized in field of content based image/video retrieval. the author in (37) addressed object detection while comparing between several color descriptors in images and videos. The histogram is the simplest descriptor to calculate, it consists of counting the number of pixel intensity values events in the image. We can distinguish several categories of histograms, we can classify them, for example according to the color space considered during the calculation: “RGB histogram”, “HVS histogram”, “Opponent histogram, associated respectively with the color spaces:“ RGB ”,“ HVS ”,“ Opponent color space ”. “The RGB histogram” is written in the normalized RGB color model ($r+g+b=1$).

2.3.2.2 Texture Descriptors

The texture is another robust low-level descriptor used for CBIR in images and videos. Several techniques have been developed to measure texture similarity. The majority of techniques compare the first-order statistics by the next one from what is known that are calculated from the images in question. These strategies work out picture surface estimations as the level of difference, coarseness, directivity, and perfection (38); or periodicity, directivity and irregularity (39). Other surface examination strategies for observing pictures incorporate the utilization of Gabo (40) channels and (41) fractals.

2.3.2.3 Shape Descriptors

Shape descriptors make it possible to present relevant data on the content of the video or image and precisely on the shape. There are many shape descriptors that differ in their simplicity/complexity. There are several shape descriptors like: CSS (Curvative Scale Space

descriptors) (42), convolution filters (43), Fourier descriptors (44), moments of Hu and Zernike (45). We are not going to go into this type of descriptor in detail because we have not used them in our work.

2.3.2.4 Descriptors Based on Points of Interest

Extracting visual descriptors from the entire image (global descriptors) minimizes the number of calculations required, the size of the dataset, and the cost of searching the most similar images. However, the holistic approach does not allow an efficient search for objects in the image. Conversely, descriptors extracted from part of the image (local descriptors) are effective in contrast to being expensive. Local descriptors can represent parts of the image obtained by applying segmentation to the entire image (by searching for regions of interest) or by searching for points of interest. Points of interest in an image are the points that will be found to be appropriate as similar images. One way to determine them is to take into account the areas where the signal is changing.

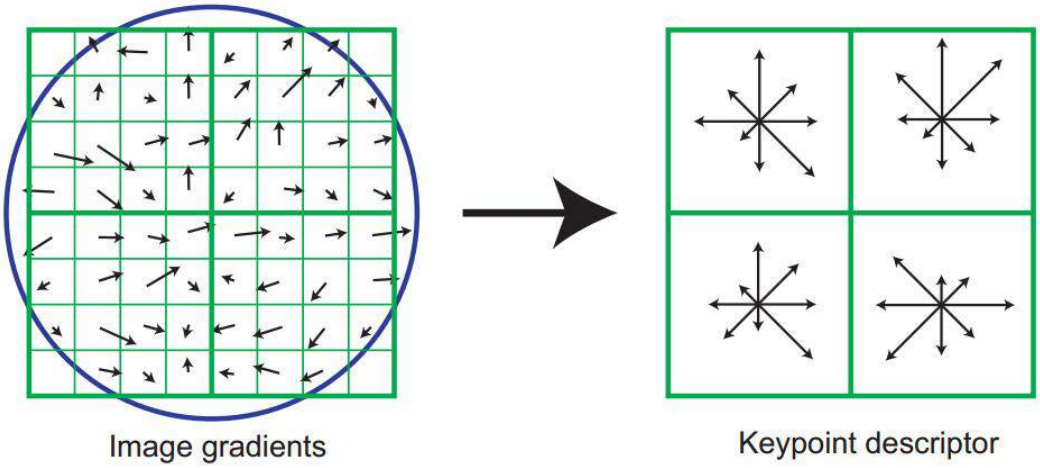


Figure 2.1: Principles of SIFT descriptors calculation. Image taken from (46)

2.3.3 Optimization of Descriptors

Before using a descriptor in a learning or classification system, it is recommended to submit it to a chain of preprocessing, this phase is called: "optimization of descriptors". Optimizing descriptors involves two important steps: descriptors normalization and dimensionality reduction. The descriptors normalization tends to modify the values of the different components of the characteristic vector. Dimensionality reduction tends to reduce the number of components forming the descriptor. An optimized descriptor tends to be more efficient in a classification approach. It is also possible to apply only one of the two methods before using the descriptor (i.e., either normalize the descriptor or reduce its dimensionality).

2.3.3.1 Normalization of Descriptors

We can distinguish two categories of normalization methods: normalization against amplitude and normalization against scale. The first category of approach is to standardize the distribution of values so that there is not a large gap between the different values. This is very useful in reducing the influence of large values that dominate small values. The second category tends to spread out the set of values so that they cover the maximum possible of a given interval. In other words, these approaches cause the nonzero part to spread over the entire scale or interval. This is similar to stretching a histogram.

X is a set of N characteristic dimensions of a given dataset that we intend to normalize where each vector x_i is composed of d dimensions: $x_i = (v_1; v_2; ::::; v_d)$. We describe on the following some normalization techniques, which are often used for the representation of images and videos.

- **L1 and L2 normalization:**

$$x'_{ij} = \frac{x_{ij}}{\|x_i\|},\ where\ : j = 1, \dots d \tag{2.1}$$

  Where x_{ij} represent the value of the j^th element of the vector x_i, and $||.||$ correspond to the norm of a vector, note that $L1 = \sum x_{ij}$ and $L2 = \sum x_{ij}^2$.

- **Min-Max:**

$$x'_{ij} = l + \frac{(u - l) \times (x_{ij} - min_j)}{max_j - min_j} \tag{2.2}$$

  x_{ij} is the j^{th} element of x_i, the minimum and maximum values of the j^{th} element in X are min_j and max_j respectively, (the i component of the different vectors $x_i \in X$). u and l correspond to the ends of the new space (target range). Usually the normalization is done so as to project the values of the resulting vector x_i in the interval [0, 1].

- σ_{norm} **:** is to center-normalize the values of each bin.

$$\sigma_j = \frac{\sum_{i=1}^{N} (x_{ij} - \overline{x}_{ij})^2}{N}\ j = 1, \dots d \tag{2.3}$$

$$x'_{ij} = \frac{x_{ij} - \overline{x}_j}{\sigma_j}\ \ i = 1, \dots n\ and\ j = 1, \dots d \tag{2.4}$$

  d represents the size of x_i, n is the number of vectors (images) of each class; the values of the mean and variance of the i^{th} bin of the vector, are $\overline{x}_i$ and σ_i respectively.

2.3.4 Dimensionality Reduction

Dimensional reduction aims to minimize the size of data by projecting it into another lower-dimensional space, without discarding significant information, more precise by keeping the maximum possible information, because this projection will cause a loss of data, that depends on the number and choice of dimensions. The objective of dimensionality reduction is to find from a combination (linear or non) of the initial dimensions of the vector in question a new space of significantly lower dimension which contains a wide segment of the total data, with the aim of finding a discriminative representation of the data. On the one hand, it helps to overcome the scourge of dimension, and on the other hand, it also helps to disrupt the data. We present in the following some approaches which are widely used to reduce the dimensionality of descriptors in the multimedia indexing system.

2.3.4.1 Principal Component Analysis (PCA)

PCA is an orthogonal linear conversion that projects information into a space of less than (or, equal) to the dimensions of uncorrelated attributes called principal components (PC), hence the name : “Principal component analysis”. It uses variance as a measure of obtained data and derives new information in such a way as to maintain as much information as possible.

2.3.4.2 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) was presented by creators in (47) as a strategy for blind source division and utilized by the creators in (48) that offer it the chance to begin acquiring truly necessary consideration in numerous spaces of sign handling, and all the more especially in unearthly examination of the picture for assignments, for example, order (49), dimensionality decrease (50), phantom deconvolution (51) or target recognition (52). The objective of this method is to find a direct portrayal of the non-Gaussian information to such an extent that the parts are pretty much as free as could be expected.

2.3.4.3 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a strategy for finding a straight blend of factors that better isolates at least two classes. LDA isn’t an order calculation, in spite of the fact that it utilizes class names. Be that as it may, the aftereffect of LDA is principally utilized as a component of a straight classifier. Then again, an elective use gets a decrease aspect prior to utilizing nonlinear order calculations. LDA dimensionality decrease frequently works on the exhibition of classifiers. In this specific circumstance, the creators in (53) showed that the presentation of their order framework has better precision after information examination by LDA than PCA.

2.4 Image Segmentation

In order to automatically locate objects in the images, the researchers first proposed segmenting the images into several regions. In general, segmenting an image consists of delimiting in the image regular or coherent ranges, that is to say, regions in which the information in the image follows an organization model: zones relatively homogeneous in intensity, in texture or color, relatively flat, smooth areas (of slow variations), etc. Many image interpretation systems are based on the segmentation of images into regions, which makes it possible to extract elementary constituents, or primitives, which will serve as a basis for the identification or reconstruction of the images (54).

Many image segmentation techniques are proposed in the literature. Segmentation algorithms are generally based on low-level characteristics. These strategies function admirably for pictures containing just homogeneous spaces of shading, for example, direct arrangement techniques in shading space as shown by work in (55, 56, 57).

We have three types of segmentation techniques, in the following we will describe each one of them [1].

2.4.1 Segmentation Based on Regions

The decay/combination type calculations exploit the particular attributes of every locale (surface, light force, colorimetry, surface, and so forth) We search for sets of applicant districts for a combination and we rate them as indicated by the effect that this combination would have on the outward presentation of the picture. The best-evaluated sets of districts are then consolidated, and it is rehashed until the qualities of the picture meet a predefined condition: number of locales, radiance, difference, or surface, or until the best scores appointed to sets of areas arrive at a specific edge.

2.4.2 Contour-Based Segmentation

This methodology tries to take advantage of the way that there is a recognizable change between two related districts. More seasoned strategies use picture handling administrators, like the Watchful channel, to feature pixels that seem to have a place with an edge. We can likewise utilize deformable models utilizing parametric bends (Bézier curve, spline …and so on) or polygons (for instance bubble calculation).

[1] https://en.wikipedia.org/wiki/Image_segmentation

2.4.3 Segmentation Based on Classification or Thresholding

The objective of thresholding is to segment an image into several classes via histograms. It is therefore assumed that the information associated with the image alone allows segmentation, i.e., that a class is characterized by its distribution of gray levels. Each peak in the histogram is associated with a class. There are many methods of thresholding. Most of these methods work correctly if the histogram actually contains separate peaks. In addition, these methods address particular case of segmentation into two classes (i.e., switching to a binary image) and their generality in the face of multi-class cases is only very rarely guaranteed.

2.4.4 Semantic Segmentation

Image segmentation is a long-standing computer vision task that consists of dividing an image into several parts, normally following some kind of criteria, but not necessarily pursuing an interpretation of image content. On another way, semantic segmentation is a different task that goes a step further than image segmentation by trying to divide an image into semantically meaningful parts (note that semantics is a branch of linguistics concerned with the meaning) (58). More specifically, semantic segmentation is concerned with dividing the image into regions with different meanings or belonging to different categories. Very often these categories are a predefined set of objects that allow total segmentation of the image. An example of semantic segmentation is depicted in Figure 2.2.

Figure 2.2: Example of semantic segmentation

A very common way to achieve Semantic Segmentation (and the default approach in recent literature) is annotating each pixel of an image according to the object they belong to. This will also be the approach pursued in this master thesis. Many consider the ability to perform semantic segmentation, sometimes also referred to as scene understanding or

pixel-wise classification, as a core capability to-wards the development of technologies such as self-driving vehicles, natural human-computer interaction, or virtual reality.

2.4.5 Semantic Segmentation Metrics

Different evaluation metrics for semantic segmentation can display different results since it is unclear how to define a good performance on this task. Three of the most commonly used metrics, and the ones taken into consideration in this master thesis, are pixel accuracy, the mean intersection over union and the mean per class accuracy. For all of them, n_{ij} is pixels number of a given class i that is seemed to belong to class j. Also, let $k_i = \sum_j = n_{ij}$ be the total number of pixel belonging to class i. If we assume to have a T which is the total number of classes, then:

- Pixel accuracy can be computed as:

$$acc = \frac{\sum_i n_{ii}}{\sum_i k_i} \tag{2.5}$$

- Mean intersection over union can be computed as:

$$miou = \frac{1}{T} \sum_i \frac{n_{ii}}{(k_i + \sum_j n_{ji} - n_{ii})} \tag{2.6}$$

- Mean per class accuracy can be computed as:

$$macc = \frac{1}{T} \sum_i \frac{n_{ii}}{\sum_i k_i} \tag{2.7}$$

2.5 Image Retrieval

Text based information are extremely restricted for data portrayal. In this manner, mixed media innovation, principle pictures, are progressively utilized. Picture dataset now addresses a huge volume of data. Questioning picture datasets is turning into a major test in the software engineering world. To successfully oversee and utilize these picture datasets, a picture recovery framework is required. This is the reason it is an extremely dynamic region for quite a while. A picture recovery framework oversees admittance to a few pictures, the pictures are addressed and ordered by the pre-owned methodology and contrasted utilizing different methodologies with produced inquiries utilizing explicit procedures. In outline, a picture recovery framework deals with the picture portrayal, inquiry development, and the determination of pictures.

2.5.1 Image Retrieval Techniques

The principal picture recovery frameworks depended on a printed portrayal of the picture (watchwords related to pictures) (59, 60, 61), catchphrases are an outside wellspring of data physically appended to pictures. The ordering system dependent on catchphrases is tedious since it is manual, precarious on the grounds that the nature of the recovery framework relies to a great extent upon the significance of the terms relegated to pictures and not normalized in light of the fact that the wellspring of these watchwords is by and large not special and relies upon the individual who partners watchwords to pictures. Furthermore, a waitlist of watchwords can't totally cover the frequently rich semantics conveyed by a picture.

The second era of picture recovery frameworks depends on content. The substance based methodology plans to straightforwardly extricate data from the actual picture to have the option to characterize it, this comes as an option in contrast to the text based methodology yet semantic issues identified with the programmed handling of pictures are immediately noted. Methods for taking care of these issues have been proposed in the writing (62, 63, 64). We can order the picture recovery approaches into two principle flows: the current of content-based picture recovery and the current of semantic picture recovery dependent on an information portrayal formalism.

2.5.1.1 Content-Based Image Retrieval

The standards of CBIR approaches are to address two inquiries (59): how to numerically depict a picture? what's more, how to gauge the likeness dependent on the theoretical depiction? These methodologies for the most part apply measurements and AI strategies. To portray pictures, most CBIR frameworks use attributes naturally separated from pictures like tone, surface, and shape. These qualities are utilized in various cycles, like similitude calculation, model structure, or even explanations.

One of the most known issues of content-based methodologies (65) is the hole between the extricated attributes from the pictures to depict it which are low-level qualities (portrayal), and the human depiction of the (picture semantic) which is of undeniable level. All in all, the hole between the picture depiction assembled consequently, and the human translation of the picture. Without a doubt, the human uses more data and qualities than the mechanized cycles to decipher the picture. Creators talk about this issue in(12). They review and portray the various strategies proposed which are characterized in five significant categories(12):

- **Using an ontology:** to characterize significant level articles and better decipher the extricated qualities of low level.

- **Using an AI strategy :** to foresee from an information measure the worth of a result proportion of an interaction or depict the association of the info measure.
- **Using a client feedback:**to attempt to get familiar with the client's aims to all the more likely comprehend their necessities utilizing web based handling.
- **Generating a semantic template:**to support significant level picture recovery by utilizing formats to address ideas determined from an assortment of reference pictures.
- **Using a web context:** for the web picture recovery. It comprises of the use of data that can be organized as the URL or HTML pages to work on the semantics of pictures.

2.5.1.2 Semantic-Based Image Retrieval

In traditional picture recovery draws near, pictures are portrayed with a bunch of watchwords, (59, 60, 66). The nature of this cycle relies to a great extent upon the significance of these watchwords, the catchphrases utilized in the client question, and the ampleness between them. For the most part, this cycle plays out a basic linguistic examination between explanation watchwords and client question catchphrases, and on the off chance that the client doesn't utilize similar catchphrases, he may not get what he is searching for. Among the issues of these methodologies is that the watchwords may not cover the rich semantics conveyed by the picture.

Semantic picture recovery approaches utilize commonly a formalism of information portrayal like depiction rationales or semantic organizations (67). The objective is to track down a model of picture portrayal. This portrayal haves to be effectively similar and describe the picture in the most ideal manner. Semantic recovery depends on the significance of catchphrases (an idea for this situation), not their grammar. The correlation between ideas (for the most part alluded to as thinking) depends on further developed procedures than those of syntactic recovery. It considers additionally the communication between various ideas of the question to further develop significance contrasted with a grammatical recovery. The creators in (68) propose an examination between metaphysics based picture recovery approaches and catchphrases based picture recovery draws near. The watchword based picture recovery is straightforward and simple to apply with OK accuracy, and the cosmology based picture recovery further develops accuracy however requires a total portrayal of pictures.

2.5.2 Facets of Image Retrieval

To plan a picture recovery framework, we want to respond to three key inquiries: how to fabricate the question? How would we address a picture? Also, how to coordinate between the inquiry and the picture portrayal (depiction)?

These three inquiries are profoundly connected to one another and are considered as the three establishments of a picture recovery framework. In the following areas of our record, we will have a concise outline of the potential responses to the above questions. The first and third inquiries concern the question modalities and likeness registering individually. We are keen on this work to a semantic-based recovery, question two is treated in a semantic-based procedures view, which as a rule addresses the picture by explanations. Different methodologies, especially on account of CBIR frameworks utilize different strategies won't be referenced here.

2.5.2.1 Query Modalities

A significant variable in a picture recovery framework is the inquiry methodology. This methodology characterizes the question language upheld by the framework as far as expressivity of created client inquiry. The most known question modalities of picture recovery framework are (59):

- **Keywords:** this methodology is the most utilized. The inquiry is introduced as watchwords. The arrangement of approved catchphrases isn't restricted overall like the picture recovery frameworks on the web however can be predefined, which will restrict the expressivity of questions.

- **Free-Text:** for this methodology, the client attempts to characterize his need by sentences. These sentences are developed uninhibitedly, it very well might be questions, stories, or different articulations.

- **Image:** like its name show, the presented question by the client is a picture and the framework returns all pictures like the client picture. This methodology is entirely appropriate for CBIR frameworks.

- **Graphics:** this methodology depends on a graphical portrayal of the question, the client characterizes graphically his requirements by drawing a picture. The picture can likewise be consequently created by a PC.

- **Composite:** it comprises of a mix of different modalities. By and large, the client can pick a methodology to characterize his question or characterize a piece of his inquiry by utilizing a few modalities. The mix of modalities is fascinating on account of intelligent questioning, the client can each time give more insights regarding his inquiry by changing the methodology. Have the effect between a picture recovery procedure dependent on watchwords and inquiry methodology dependent on catchphrases. Question modalities are utilized as a contribution to the inquiry development process, which gives

after, a question to the picture recovery process, paying little mind to the preowned procedure.

As a rule, the question development process isn't required for syntactic recovery procedures dependent on watchwords, the arrangement of catchphrases presented by the client are utilized without pre-treatment. By cons, it is vital for semantic picture recovery methods to give some semantics to the question.

2.5.2.2 Images Annotation

It will be described In section 2.6.(i.e., Image annotation).

2.5.2.3 Images Similarity

Similitude registering is an interaction used to recover the arrangement of picture replies. It characterizes a fondness relationship that can be measured or not between a question and a picture. The inquiry is produced utilizing the yield of the question methodology, it could be another picture, text, or a mix of both. The comparability figuring likewise permits commonly the positioning of picture answers regardless of whether different methodologies utilize different boundaries, also, to do the positioning (69, 70). The likeness can be syntactic, semantic, or half and half. The syntactic likeness depends on a mathematical correlation between the portrayals of the visual parts of the pictures and the inquiry. The consequence of this correlation is mathematical and permits the positioning of the appropriate responses. By cons, semantic similitude depends on the correlation between the understanding of the portrayals of visual parts of pictures and the translation of the inquiry, it by and large uses an information portrayal formalism and an information base.

This information base is utilized to all the more likely decipher the portrayals of the pictures and the inquiry for a superior correlation. The aftereffect of the correlation is for the most part not numeric however boolean and doesn't permit essentially the positioning of picture replies. The mixture likeness utilizes a blend of syntactic and semantic similitudes. It utilizes the information base to improve and for a superior correlation, and numeric calculation for the positioning.

A. Syntactic similarity

The general principle of syntactic similarity is the computing of numerical distances between the query and potential answers. The distance is defined according to the used model, it may be for example one or a set of values, numerically calculated using the representation of the query and the representation of a potential answer. The selection of answers is based on the values of distance, for example, by defining a distance threshold or the selection of N

first answers after the ranking. We can distinguish two main approaches for computing the syntactic similarity:

- **Visual signature using:** These methodologies are fundamentally utilized in CBIR frameworks. The standard is to numerically plan the qualities called marks. These marks are extricated from the pictures for utilizing the methods of distance figuring to observe one or a bunch of qualities addressing the likeness.
- **Information recovery models:** The guideline is basic, it comprise to utilize text in type of watchwords, labels, sections, or different structures to file pictures. The thought is to change picture recovery to data recovery to utilize methods and models of data recovery.

B. Semantic likeness

The likeness, for this situation, isn't a distance to figure typically yet a course of contrasting a component of a given language. An information portrayal formalism is utilized to officially characterize an-documentation picture and a question. This formalism has distinct semantics for deciphering the explanations and inquiries. The administration of understanding is called thinking. The thinking is utilized to look at an explanation and an inquiry. The correlation is normally sensible and in light of set hypothesis, it can utilize standard or non-standard thinking.

- **Standard reasoning:** These are traditional normalized thinking which has a place with the standard meaning of the language. This thinking is for the most part dominated, essential for the best abuse of language, and remembers for its specialized devices, for example, the device which permits thinking called reason-er. They are additionally the reason for characterizing other thinking called nonstandard.
- **Non-standard reasoning:** They comprise of non-traditional thinking utilizing mathematical processing or likelihood hypothesis frequently. They are frequently characterized for explicit applications and not accessible on specialized instruments like reasoners. This thinking is for the most part dependent on standard thinking however can likewise utilize non-standard semantics.

2.5.3 General Architecture of an Indexing System

Typically, a visual content image search system has an offline image base indexing phase and an online search phase. Figure 2.3 represents the general architecture of an indexing system and content-based image retrieval, this system is executed with two stages: the indexing stage and the search stage.

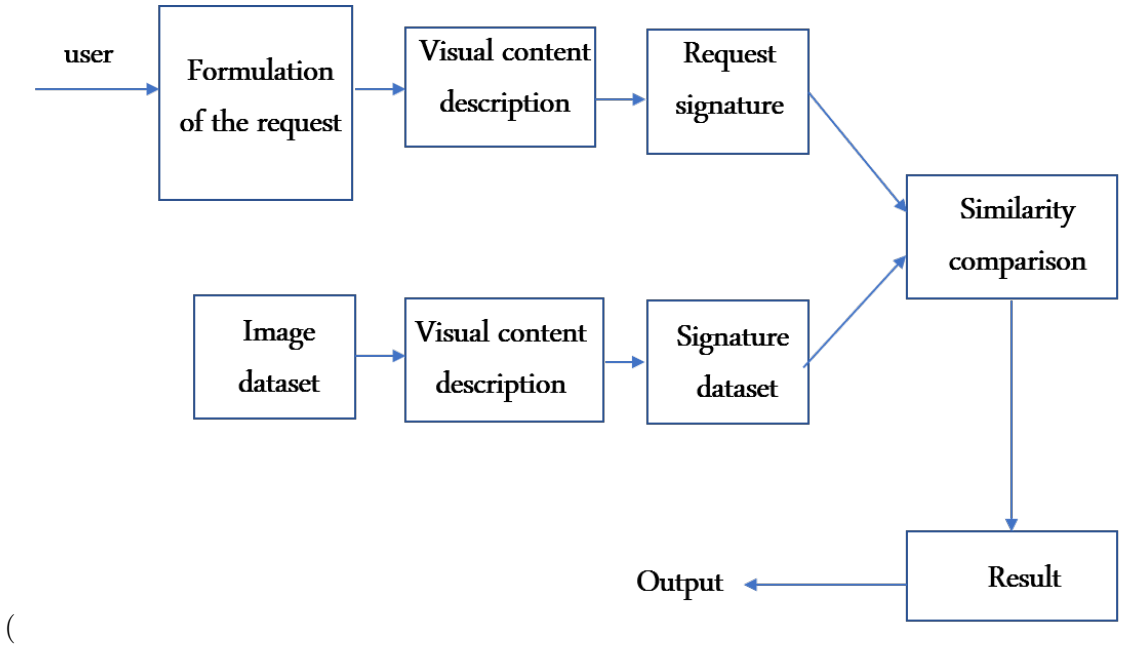
- The purpose of the indexing step is to organize and prepare the dataset. This phase is said to be offline because it takes place before any research. It includes the following treatments:
  - Extraction of characteristic descriptors from images.
  - The construction of indexes from descriptors. The goal is to put in place the techniques to access any descriptor as quickly as possible during the search.
- The search step takes a query vector as input and uses an algorithm that takes full advantage of the indexing step, thus confining the search for similar data. In other words; while browsing the dataset, the user chooses an image through a graphical interface. The indexes or signatures of the request are compared to the indexes of the reference images, so the system selects and presents to the user the images that are most similar to the request. The visual content of the dataset of images that are extracted and described is called image signatures. The signatures of the images in the dataset constitute a dataset of signatures. To search for images, the user provides a sample image (called a query). Different methods of formulating requests exist, the treatment of these in the system depends on the way in which the information is presented to it. A search model specifies how the query is represented. Below we list the types of image querying methods that can be classified into six categories (71):
  - **Query on a single characteristic:** the selection of images is based on a single characteristic, the percentages of different values taken are determined by the user. An example of this type of query is to find images that contain 10% red, 30% green, and 60% blue.
  - **Simultaneous request on several characteristics:** the user request is a combination of several characteristics (color, texture, shape). An example of this type of query is to find images containing 10 % red and 30 % green and 60 % blue of the tree texture.
  - **Request on the location of the characteristics:** the user specifies the different values taken by the characteristics and the location of these characteristics in the image. An example of this type of query is to find images where 25% of the color red is located to the left of the image.
  - **Query by sketch:** the user draws his query using a graphical interface to find images that look like him in the dataset. There are two types of drawing

    - **The sketch (72):** the user describes what he wants by precisely representing the outlines of the objects, usually in a single color, only the shape aspect carries the information.
    - **Rough drawing (73):** a colored drawing is suggested by the user. It contains a colorful representation of each object, but or (the "maize" makes the sentence heavy) the outlines are generally vague. Color information (the colors themselves but also the arrangement of them) is therefore essential, unlike the shape which is not very representative.

- **Search for images by objects:** the user describes the characteristics of an object in an image rather than the entire image, the goal is to search for a specific object in a series of images. An example of this type of query is to find images containing a person. You can also combine several objects and specify the spatial relationships between the different objects. An example of this type of questioning is finding the pictures where a person is sitting next to a tree.
- **Search by example:** In this case, the system needs to compare an example of the same type (image) with the dataset to produce similar images. This method is simply natural and does not require extensive knowledge to manipulate the system. It is therefore well suited to a non-specialist user. After formulation, the system represents the request by its signature. The measures of similarities/dissimilarities between the signature of the request and that of all the images in the dataset are calculated and compared. The result is most often presented as a list of images of descending similarity. As all the images in the dataset must be examined for images similar to the query, the cost becomes prohibitive as the size of the dataset increases. To remedy this problem, most CBIR systems use an indexing scheme that provides a very efficient search method. The main idea is not to browse all of the dataset but to examine a single small part of the dataset (74).

2.6 Image Annotation

The description of the images is done using annotation techniques. This section will be devoted to these annotation techniques. Three annotation techniques are proposed in the literature: manual annotation, automatic annotation, and semi-automatic annotation. Picture explanation is a picture portrayal process that works with admittance to pictures. It is essentially utilized by semantic-based picture recovery procedures however can be likewise utilized by syntactic strategies. The thing that matters is that the semantic methods utilize



(

Figure 2.3: Diagram for a typical CBIR

an information portrayal formalism to characterize an explanation, and an information base to enhance and to more readily decipher the comments.

To portray a picture, we can partner to it a few kinds of data. This data are (60, 75):

1. **content-free meta-information:** this data doesn't portray the substance of pictures. They depict the setting as the creator of the picture, the pre-owned camera, the picture goal, and so on

2. **visual content meta-data:** this data depicts the visual substance of pictures. It tends to be isolated into two kinds:

   (a) **content-subordinate meta-data:** depicts the low level attributes of pictures. By and large data can be naturally removed like tone, shape, surface, and so on;

   (b) **content-graphic meta-data:** depicts the semantic of pictures content. This data addresses the potential translations that a man can provide for pictures.

2.6.1 Manual Annotation

Manual image annotation consists of annotating a dataset of images by one or more annotators. During manual annotation, the annotator assigns a description to each image according to its perception. Several manual image annotation framework have been presented in the literature (76, 77, 78). Keyword-based image annotation is a process of describing images

with terms in order to facilitate retrieval. The authors in (79, 80) provided an exhaustive study on the contours and complexity of annotations.

A keyword is a word associated with an image that, once indexed, identifies the image in a dataset. It allows making a textual description of the content of the images. Like all other descriptors, keywords are used on the one hand to describe images, and from another meaning to access images through textual queries. Compared to low-level descriptors (color, texture, shape, etc.), keywords are considered effective descriptors. With keywords, annotation, retrieval, and even comparison of images become possible. Several approaches have focused on the description and retrieval of images via keywords. The first frameworks showing up in quite a while depend on the text (23, 81, 82). They adopt the strategy of portraying visual substance in literary structure. The catchphrases utilized fill in as a file to get to the related visual information. The benefit of this methodology is that it permits you to look datasets utilizing standard question dialects, for instance, SQL. A few methodologies depend on the depiction of the visual substance of pictures ((83, 84, 85, 86).

Another method is to combine visual content and textual content. Unlike visual content described by keywords, textual content represents all the attributes specific to images, namely: titles, captions, comments, etc. The combination of the visual and textual content has been proposed by several researchers (87, 88, 89) and performed image annotation using a consistent language model because the keywords alone would be independent. For them, there is a need to take into account other information such as the text surrounding the images, in order to produce a good description of the image. We also have the multilevel annotations offered in(88) which not just plan to recognize explicit items in a picture, yet in addition fuse the ideas to gather comparative components together.

Usually, the process of manual annotation involves the annotators who describe the image, the image to be described, and the descriptors provided to complete the description as shown in Figure 2.4.

2.6.2 Automatic Annotation

Automatic annotation is a strategy of annotating images through an annotation system. The automatic annotation process usually takes place when one or more new images are added to the dataset (90, 91). The system automatically uses each new image as a query and searches for images by content. For a number n of images similar to a query, the characteristics used for the annotation of each of them (and classified by their frequency) are analyzed. A list of characteristics of images that are identical to the new image is provided and assigned to the new image. The new image is thus annotated (although virtually and without confirmation). Several works have already been carried out in the field of automatic annotation. Among

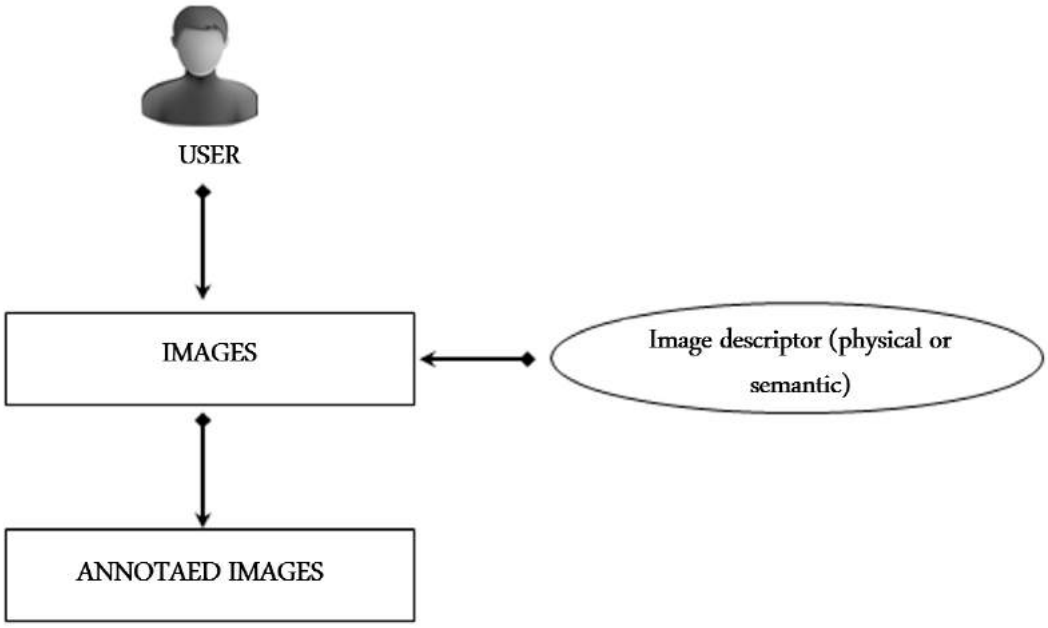


Figure 2.4: Manual annotation process with elements of the image, such as a computer screen and mouse pad, but also the desktop and the interior of the office

these works, we can cite the methods proposed in (90, 90, 91, 92, 93). Figure 2.5 shows the automatic annotation process. Annotations for new images will be made from this sample. The following sections describe the methods for automatically annotating images.

- **The method based on classification:** it allows to create a very wide number of classes to which the images are associated (94, 95, 96, 97). It allows the description of the characteristics of images in the form of vectors ((98)). The images are represented by vectors of their characteristics. Graphs are often used to classify images. The vectors and the keywords of the images are represented as nodes of a graph $G = (V, E)$, note that V is the group of nodes. And E the group of edges as the confirms the work carried out in (99, 100). Most CBIRs are built on the vector space model. The image and query corpus are descrived as feature vectors in an n-dimensional vector space

Typically, the picture is first investigated by highlight vector extraction, and the words for comment preparing are utilized with AI strategies to have the option to consequently relegate comments to new pictures. The initially was to think about the connections between's the descriptors of pictures and the explanations, after that, new techniques appeared that use methods resulting from the machine translation emerged, allowing to make a translation between the textual vocabulary, and what one considered a visual vocabulary, where each word would be a group of similar regions, obtained by clustering. Classification is based

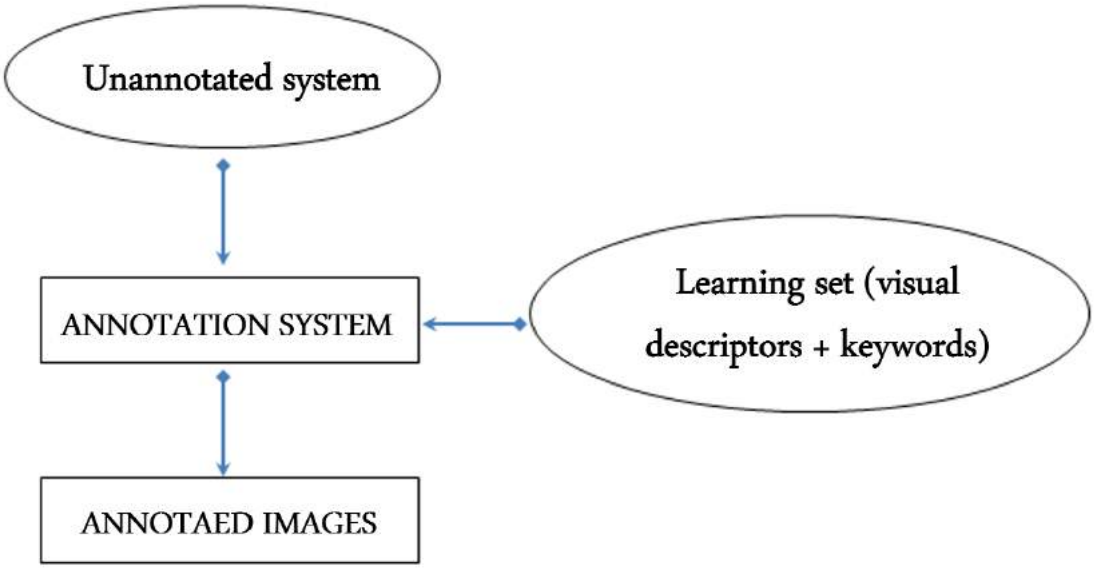


Figure 2.5: Automatic annotation process

on learning. Learning is a process of improving the performance of a system based on past experiences. Three main families of learning techniques are listed (101):

- **Supervised learning**: is a type of a ML strategy that automatically generates a classification based on existing examples. Supervised learning is possible with some classifiers. This is the case, for example, with SVM Support Vector Machines (102), random forests (103), the k nearest neighbors algorithm (denoted KPPV, or KNN in English), and Bayesian classifiers.

  SVMs and random forests are known to perform well in the presence of a wide number of variables. On another way, the use of SVMs becomes difficult to use when the number of observations of the learning base is large. Regarding KNNs, the classification procedure is cumbersome because each requested image is compared (on the basis of its characteristics) to all the stored images. On the other hand, this method has the advantage of not requiring any training: it is the sample that constitutes the model.

  At last, Bayesian classifiers, then again, are delicate to the dimensionality of information. Be that as it may, they are compelling with a ton of learning information. Bayesian classifiers are appropriate for tackling issues within the sight of missing information, not at all like SVM (104). Other learning strategies, for example, neural organizations are additionally utilized for idea learning. In (105), the creators pick 11 classes of ideas: block, cloud, hide, grass, ice, street, rock, sand, skin, tree, and water. Then, at that point, a lot of preparing information is taken care of into the neural

organization classifier to relate the low level elements of the picture to its semantics (classification of names).

- **Unsupervised learning (sometimes referred to as "clustering")**; is a method of ML. It is for a strategy that makes it possible to divide a heterogeneous group of data, into subgroups so that the data considered as the most similar are associated within a homogeneous group and that, on the contrary, the data considered as the most similar. data considered to be different are found in other distinct groups. During the unsupervised learning phase, the system must determine its outputs based on the similarities detected between the different inputs (self-organization rule) (106). In contrast to directed learning, solo learning unites a bunch of picture information in a manner that expands similitude inside bunches and limits likeness between various gatherings. Each subsequent gathering is related with a class mark and the pictures in a similar group are thought to be like one another.

- **Reinforcement learning:** This is learning that requires a supervisor to tell the agent what action is correct in a given situation (107). In reinforcement learning, the agent interacts with the environment which gives him quantitative feedback on the values of his actions. The objective of reinforcement learning is then to generate from experiences (current state, action, next state) a policy maximizing performance over a given period.

- **The probabilistic method:** it allows images to be represented by an n-dimensional characteristic vector. These vectors are calculated using mathematical functions or algorithms. A good choice of relevant characteristics is essential to achieve great discriminating power (98). The probabilistic method is often dedicated to the annotation of partially annotated images, that is, pictures that don't have the most extreme number of catchphrases. At the point when a picture is to some extent commented on, missing watchwords are viewed as missing qualities. The probabilistic technique is a strategy that permits computing the dissemination of the watchwords to a picture (108, 109).

  This circulation addresses a forecast of missing catchphrases from a picture. For each missing annotation, the keyword of the vocabulary with the greatest probability is retained, if this probability reaches a certain threshold (110). Each image is labeled with the keywords with the most noteworthy probability. This is the case of (87) which provides a language model for annotating images that calculate the probability of a group of keywords. The group of keywords with the most noteworthy probability is associated with the image if that probability exceeds a certain threshold. Other work on the automatic annotation of web images has been proposed by some researchers. This is the example of works in (111, 112). It is a method that makes it more straightforward

to observe pictures because of extra data accessible on the web, for example, model the URL of the picture document which regularly has an unmistakable progressive construction, including data about the picture like the classification of the picture.

Clients need to peruse the whole rundown to track down the ideal pictures. This is an extended interaction in light of the fact that the returned results are assortments of pictures. To work on the exhibition of these methodologies, the scientists endeavored to consolidate the printed data and the visual substance of the picture. This is the point of the work completed in (112) which consolidates literary portrayals (HTML) and visual qualities of the picture.

To do programmed comment, they created two free classifiers. The first is text-based and the second is visual-based. The analysis was completed on a predefined set of 15 ideas. The outcomes got show the significant exhibition of the framework. Notwithstanding, because of the imprecision of extricating text based data, the exhibition of certain ideas isn't good.

Different proposition have been made in (113, 114). Microsoft Research Asia (MSRA) plans to total query items from exemplary web picture web indexes with the goal that clients find pictures rapidly. Initial, a savvy vision-based division calculation is intended to section a page into blocks. In the square containing the picture, literary data and connection data of a picture can be extricated exactly. Then, at that point, a diagram of the picture is made utilizing strategies for examining joins between blocks. Consequently, for each picture, we acquire three kinds of portrayals, the visual portrayal dependent on the actual attributes, the semantic portrayal dependent on the printed qualities, and the realistic portrayal. The pictures in every class are then improved by their visual attributes. What a large portion of these strategies share practically speaking is that they require a preparation test.

2.6.3 Semi-Automatic Annotation

The semi-automatic annotation results from the combination of the two preceding annotations. It consists of generating the annotation of the images from the manual annotation and the automatic annotation. (113). In general, semi-automatic methods consist of involving the user to validate the decisions of the system (115, 116). This collaboration can be done in both directions: either it is the human who verifies and validates the annotations made by the machine, or either, it is the machine that completes the annotation process initiated by the human. Semi-automatic annotation techniques are generally based on relevant feedback. A variety of user interfaces for image search and relevance feedback are used for the semi-automatic annotation method (91). Typically, such a user interface consists of three parts: the query submission interface (either a keyword query, or an image query, or a combination of both), the image browser, and the relevance feedback interface.

A typical user scenario is as follows: When a user submits a query, the system returns search results as a ranking list of images based on their similarity to the submitted query. Images with higher similarity have higher ranks than those with low similarity. The image browser can be a scrolling window of images, a paginated window. The user can browse the images in the browser and use the confirmation interface to submit their relevant judgments. The system iteratively returns refined recovery results based on user feedback and displays the results in the browser. This process is illustrated in Figure 2.6. The relevance feedback allows the most relevant images to be displayed in the first rows and gives the user a better chance to see them, to confirm them, and therefore to annotate them.

There are two cases to consider at this stage (91, 116). In the first case, there are no images and in the second case, some images are already annotated with the keywords corresponding to the request. In the first case, the system only returns a list of random images because no keyword is matched and no image relevant to this query can be found. In the second case, the images annotated with the keywords of the query are retrieved and displayed to the user. It is then up to the user to indicate which images are relevant. For each of these relevant images, if the image has not yet been annotated with all the keywords of the query, the image is annotated with the keywords with an initial weight of 1. If the image has already annotated with one of the keywords present in the query, the weight of this keyword for this image is increased by 1. For each of the irrelevant images, the weight of this keyword is reduced by a quarter $\frac{1}{4}$ of its original weight. If the weight becomes very low (e.g., less than 1), the keyword is removed from the image annotation. The result is a set of keywords and their weights associated with each image and stored in a dataset.

2.6.4 Other Annotation Approaches

As we have just observed, each of these three techniques presented above is carried out individually and with descriptors such as low-level characteristics and keywords. Recently, new annotation approaches have emerged. It consists of integrating semantics on the one hand and collaboration on the other hand in the annotation system. In this section, we present the collaboration-based annotation approach and the semantic-based approach.

2.6.4.1 Collaborative Approaches

Some approaches have looked at collaboration. Collaborative annotation consists of the proposal of the description of an image by several annotators as explained in the work carried out in (117). It can be achieved either by the use of keywords. Usually, several annotators annotate the same image and the system deduces the most relevant description of the image. In (118), a game is offered to Internet users. The best-known example is ESP Game.

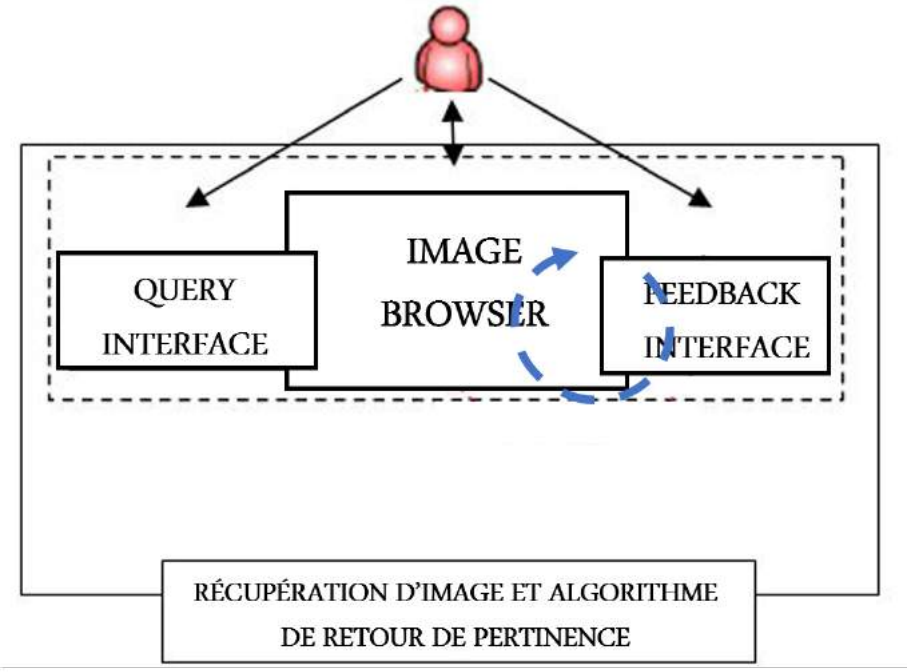


Figure 2.6: Architecture of the semi-automatic annotation in the system having been annotated with the keywords of the submitted query

The rule of the game is as follows: two users, connected at the same time, are offered by the system of the same image. Each one proposes keywords to annotate this image. The system does not offer vocabulary. Once both users agree on a keyword, that keyword is validated, and both users earn points. Another image is then offered to them, and so on. When a user reaches a certain number of points in a week, they are given a gift. Authors in (119) have proposed Image Webs. This is an approach that favors collaboration between annotators. It takes into account several annotators and allows them to share not only the annotated images but also annotations (made with keywords). Using this approach, the annotators describe the different regions of the images by keywords. The Image Webs system connects all of the image regions containing the keywords of the same object to form a graph. These graphs represent the relationships between images based on their shared visual content.

The (120) approach provides a collaborative server-based annotation system that can simply be accessed through a web browser. This helps maintain annotation statistics and avoids distributing large collections of images to users. The approach aims to simplify and speed up the annotation process as much as possible, while at the same time maintaining configuration and customization options to allow for different annotation styles and user preferences. Users will be able to use different annotation styles, such as displaying and annotating a few images per page; annotating images with a single term before switching to annotating with multiple terms simultaneously, etc. To maximize user convenience and

efficiency, each user then has the ability to customize the number, size, and arrangement of images displayed per page, and select one or more concepts to annotate at a time.

In (121) the proposed collaborative image annotation approach allows users to create tags that provide a hierarchical context to define the relationships between these concepts. The approach also provides a technique to establish the credibility of the users who annotate the images. It is a technique for calculating the veracity and reliability of a particular statement. A similar approach has been proposed by authors in (122). They provide a collaborative environment for tagging video sequences with keywords. The environment allows annotators to share their own annotations with others, thus speeding up the process of generating annotations. The environment has an interface that allows annotators to annotate video footage and share annotations with other annotators.

Authors in (123) present a new method of collaborative annotation of images. It is an approach to annotating images by a network of people. It consists of extracting metadata from conversations between groups of people talking to each other about a given image. The approach explores how language can help annotate images. The approach consists of three steps: the detection of a set of meaningful labels that can be associated with each image by the variety and richness of the language used by users when they start talking about their pictures, recognition of the semantics of these terms by the interactions between groups and the propagation of the labels. It is an approach that explores a multimodal fusion technique to recover the spelling and pronunciation of redundant terms in speech and writing.

2.7 Performance Metrics

In (124), several ways to evaluate the performance of self-annotations are provided:

- Ask novice users to evaluate the results provided by the system.
- Measure performance from keywords associated with the image, such as the words surrounding an image from the web.
- Measure performance from keywords associated with images manually by professionals. The set thus annotated is then called ground truth. This is the case with the Corel dataset.

2.7.1 Empirical and Random Score

To measure the quality of a predictive model, it is important to take into account the difficulty of the task to be performed. For example, if we want to predict the weather, and we know that the weather is good 350 days a year, a model that predicts that the weather is always

good only makes 4 % errors. To measure the quality of the results of a predictive model, it is important to measure the score obtained compared to that of a model built only on a priori knowledge (called empirical model).

Automatic annotation systems tend to predict mostly very common words, such as sky, water, people, and very few infrequent words, such as anemone, cactus, elephant. A model that annotates images with the scores of the most common words efficiently. However, this model may not in fact provide any new information. This is why it is important to compare this score with the empirical score obtained from the frequency of the words in the database (a priori distribution). In addition, this makes it possible to compare the scores of different models obtained on data with different difficulties.

The empirical score should not be confused with the random score obtained by the random classification of data. Concretely, it can be obtained by calculating an a priori distribution of the words in a random manner.

2.7.2 Measure Quality of Distribution a Posteriori

For predictive models, the authors in (96) propose to measure the quality of the posterior distribution of words by calculating the Kullback-Leibler divergence between the distribution of words $p(\omega, B^d)$ produced by the model by knowing the set B^d blobs of the image d and the actual distribution $p(\omega)$ of the words in that image. Unfortunately, this latter distribution is unknown. However, it can be assumed that the words which should actually annotate this image follow a uniform distribution and that the other words are not predicted, in other words:

$$p(\omega) = \begin{cases} \frac{1}{n_{w^d_{ref}}} \, if \, \omega \in w^d_{ref} \\ 0 \, Otherwise \end{cases} \tag{2.8}$$

By definition, the error on a d document is calculated by:

$$\begin{aligned} E^{Model}_{KL}(d) = \sum_{w \in W} p(w) log \frac{p(w)}{p(w|B^d)} \\ = \frac{1}{w^d_{Ref}} \sum_{w \in w^d_{Ref}} log \frac{p(w)}{p(w|B^d)} \\ = Constante - \frac{1}{w^d_{Ref}} \sum_{w \in W^d_{Ref}} log \, p(w|B^d) \end{aligned} \tag{2.9}$$

où $Constante = -log \, n_{w^d_{ref}}$

To measure the performance of a group of images, it suffices to calculate the average of the $E_{KL}^{Model}(d)$. To measure the performance against the empirical model on the data of the test set T, it suffices to calculate:

$$\Delta_{KL} = \frac{1}{nT} \sum_{d \in T} (E_{KL}^{Empirical} - E_{KL}^{Model}(d)) \tag{2.10}$$

Δ_{KL} is negative when the model score is lower than the empirical model score, positive otherwise. A similar measure is proposed in (110). The quality of a model's annotation is assessed using a classic measure of language processing community called caption perplexity:

$$perplexity = exp\{-\frac{\sum_{d \in T} \sum_{w \in W_{Ref}^d} log\, p(w|B^d)}{\sum_{d \in T} n_{W_{ref}^d}}\} \tag{2.11}$$

For this measure, the lower the score, the more efficient the model. We note that these two measurements only take into account the words that are in the initial caption.

2.7.3 Recall and Precision

Review and accuracy are two exemplary estimations in data recovery. The review is the proportion of the quantity of pertinent reports found to the complete number of important records. Accuracy is the proportion of the quantity of applicable records found to the complete number of chosen reports.

Let n be the quantity of important reports, r the quantity of significant records found and w the quantity of non-applicable archives found. The callback R and the accuracy P are characterized by:

$$R = \frac{T}{r}\, and\, P = \frac{r}{r+w} \tag{2.12}$$

To use these measures within the framework of the auto-annotation, we carries out for each word of the lexicon a query of $q = w$ that comprising only a word. We then count the number $n_{W_{\neq \varnothing}}$ of words for which at least one image has been found:

$$n_{W_{\neq \varnothing}} = \sum_{w \in W} |\{w|R(w) > 0\, and\, P(w) > 0\}| \tag{2.13}$$

Then we measure the average recall mR and the average precision mP on all the words of $W_{\neq \varnothing}$:

$$mR = \sum_{w \in W_{\neq \varnothing}} \frac{r(w)}{n(w)}\, and\, mP = \sum_{w \in W_{\neq \varnothing}} \frac{r(w)}{r(w) + w(w)} \tag{2.14}$$

Where $n(w)$ is the quantity of test pictures clarified with w, $r(w)$ is the number of images whose caption initially contains the word and which the system has annotated with the word, $w(w)$ is the number of images not initially annotated by this word and annotated by the system with this word. When using the measures mP and mR, it is important to specify the number $n_{W_{\neq\varnothing}}$ of words for which at least one image has been found, because a system can get a strong average recall and high average precision by predicting only the most frequent words.

2.7.4 Normalized Score (NS)

A widely used metric (96, 125) to quantify the exhibition of auto-comment frameworks is the Normalized Score (NS). We first give the general definition that is valid for any information retrieval or classification system with two classes: relevant and irrelevant elements, then we apply it in the case of auto-annotation systems based on an image.

Formally, let N the quantity of things in the set to be classed, n the quantity of significant things, and r is the quantity of established. Because of that, the overall meaning of the NS score is:

$$NS = \frac{r}{n} - \frac{w}{N-n} \tag{2.15}$$

This score is comprised of two terms: the first is the quantity of significant things found and standardized by the quantity of pertinent things (likewise called review or affectability), the second is the quantity of insignificant things found and standardized by the quantity of unessential components (it is equivalent to 1-explicitness). The score NS is between -1 and 1. $NS = 1$ when all the elements found are all the relevant elements ($r = n\, and\, w = 0$). $NS- = 1$ when all elements found are all irrelevant elements ($r = 0\, andw = Nn$) when all elements are found or no element is found ($r = n\, and\, w = Nn\, or\, r = 0\, and\, w = 0$).

- **Average normalized score)**: In the case of an auto-annotation system where each image d of the tests T has a set of $n_{W^d_{Ref}}$ relevant words, the average NS score (NS^{model}_{Apr}) is:

$$NS^{model}_{Avr} = \sum_{d\in T} \left(\frac{r(d)}{n_{W^d_{ref}}} - \frac{w(d)}{N - n_{W^d_{ref}}} \right) \tag{2.16}$$

Where $r(d)$ is the number of words in the caption of d that the system has actually associated with the image, $w(d)$ is the number of words that are not part of the caption, but which have been associated with the image by the system, N is the number of words in the lexicon. The author in (126) provides a matrix version of the average NS score. In order to

be able to compare different models, it is preferable to calculate the difference $\Delta_{NS_{Apr}}$ and the gain $G_{NS_{Apr}}$ on the empirical model:

$$\Delta_{NS_{avr}} = NS_{avr}^{model} - NS_{avr}^{empirical} \; and \; G_{NS_{avr}} = \frac{\Delta_{NS_{avr}}}{NS_{avr}^{empirical}} \tag{2.17}$$

A measure close to the NS score is used by authors in (96, 125). For each image, we assume that the system predict exactly the number of words in the caption of the image ($r + w = n_{W_{ref}^d}$), then we measure the annotation accuracy by calculating (125):

$$Acc(d) = \frac{r(d)}{n_{W_{ref}^d}} \tag{2.18}$$

The word prediction score (96) is then calculated as follows:

$$PR_{avr}^{model} = \sum_{d \in T} \frac{r(d)}{n_{W_{ref}^d}} \tag{2.19}$$

This score assumes that the number of words that annotate an image is known. However, this information is not known when one wishes to annotate new images. This measurement, therefore, uses information that can distort the actual performance of the system.

We notice that in the case where the number of words in the lexicon is very large compared to the number of predicted words ($N >> (r + w)$), the average prediction score is a good approximation of the average NS score. However, this measurement does not make it possible to compare experiments carried out on corpora containing a number of different words. For example, let a corpus A whose lexicon includes 100 words and a corpus B, which can contain the same images, whose lexicon includes 1000 words. Suppose that from the two corpora, we annotate an image d and obtain just 1 word out of 5, we will have $Acc_A(d) = Acc_B(d) = 1/5 = 0.2$, but $NS_A = 1/5 - 4/100 = 0.16$ and $NS_B = 1/3 - 4/1000 = 0.196$. The NS score takes into account the difficulty of the task (the probability of making a mistake is greater when you have a lexicon that contains a lot of words), but not the *Acc* score.

2.8 Machine Learning / Classification

Machine learning is applied in cases where a programmer cannot explicitly tell the machine what to do and what actions to take. Nowadays, and for several years, several applications benefit from machine learning. The typical example of this kind of application is one in which we need to return a list of ordered results in terms of relevance to a given query. Among the best known and most used by the majority of users, we can cite automatic translation applications, face recognition which can be part of security systems or access controls to places or services, recognition voice, fingerprint recognition, etc.

The objective of AI is to gain from a bunch of information that are known as a "learning set", a rundown of data so similar sorts of data can be gotten from information that has not yet been seen. AI can deal with a few issues, for example, overfitting, it is characterized as a demonstrating mistake that happens when the model learns a capacity to adjust to a restricted arrangement of information, and which will thusly experience issues summing it up to new examples. Truth be told, the information concentrated frequently have some level of mistake or arbitrary commotion. Accordingly endeavoring to cause the model to adjust too near a modest quantity of mistaken information can taint the model with significant blunders and diminish its prescient power (force of speculation). To estimate the reliability of the learned model, to optimize the parameters of a model and/or to avoid the problem of overfitting, a method called "cross-validation" is used, which consists in dividing the learning set into two parts: one to build the model and the second to evaluate it. There are different possible ways of operation, such as:

- Divide the training set into two parts, typically > 60 % of the samples for model training and the rest for testing. The error or a performance measure is estimated depending on the problem in question (e.g., Precision, root mean square error).
- Divide the training set into k parts, to train the model on one of the $(k-1)$ sets and evaluate it on the remaining set. The operation is repeated k times, selecting at each iteration a validation set that didn't take into account in the previous iterations. This procedure is called "K-fold cross-validation". The performance measures or errors calculated as iterations are averaged to calculate a performance measure / final error.

Cross-validation is very often used for the optimization of the parameters of the learning methods and also to avoid the problem of overfitting.

2.8.1 Supervised v.s. Unsupervised

There are two classes of AI; managed learning and solo learning. In managed learning, the student intends to choose the best capacity $g : X! Y$ permitting to coordinate with a bunch of m information x_i written in a space X to target classes or classifications $y_i \in Y$. The objective is to pick the most proper exact capacity that meets at least one improvement rules while holding the force of speculation for tests that not yet seen. In administered learning, the student has a learning set $D_{train} = (x_i, y_i)m_i = 1$ where $x_i \in X \, and \, y_i \in Y$.

In other words, each element of the learning set is annotated. The learner will therefore use the representation of each sample as well as its annotation to learn a function g and to generalize it for data that is not used i.e. data test. If the labels or classes are of continuous type (that is, real values), it is called regression.

In unsupervised learning, the classes of the learning samples are unknown. Therefore, the training set will consist only of the descriptions of the training samples: $D_{train} = (x_i \in X)\, m_i = 1$. The learner tries on the basis of certain measurements, typically distances, to study the existence of clustering of samples. The samples in each cluster are supposed to have characteristics in common. The K-means method is one of the best-known clustering methods.

There is another machine learning class that is called semi-supervised, which is sort of an intermediary type between supervised and unsupervised methods. Semi-supervised learning makes use of a set of data which is not all annotated, typically a small number of annotated data and a large amount of untagged data: $D_{train} = (x_i, y_i)\, m_0^i = 1[(x_i)\, m_i = m_0 + 1$.

In (127), the authors show that it is possible to transform a multi-class classification problem into several two-class problems using the "**one-vs-all**" principle. Each binary system classifies the samples into one class or another that includes all of the remaining classes. There is another "**one-on-one**" strategy, which is to generate one classifier for each pair of classes. The class that receives the most votes is assigned to the sample in question.

2.8.2 Generative Approaches

Given a set C of m classes and a set of samples X, the generative approaches model the joint probability $p(x, c_i)$ of a sample $x \in X$ and a class $c_i \in C$ and predict for x which class is most likely to rank to. To do this, they rely on the computation of the probabilities $P(c_i|x)$ using Bayes' theorem:

$$P(c_i|x) = \frac{P(x|c_i) \times P(c_i)}{P(x)} \tag{2.20}$$

- $P(ci|x)$: The posterior probability of c_i considering x.
- $P(c_i)$: The apriori probability of c_i, also called the marginal probability of c_i.
- $P(x)$: The apriori or marginal probability of x.
- $P(x|c_i)$: The probability of x giving c_i, with c_i is a known parameter (fixed). It is also known under the names of:
  1. Likelihood function of c_i.
  2. The density' probability of the class c_i. In the case where the apriori probabilities are equal for the different classes, the decision can be made based only on the likelihood functions $P(x|C_i)$ of each class. A typical generative method relies on a Gaussian mixing model (GMM) (128) to model the distribution of training samples.

The full arrangement of GMM boundaries can be viably picked up utilizing the ”most extreme assumption” calculation (129). Bayesian techniques are especially appropriate when the size of the information is little.The parameters are estimated using the maximum likelihood method (130). The popularity of the Bayesian classification is accentuated in the field of text analysis and searching (131), especially for SPAM detection (132), and emails classification (133). In addition to its simplicity, this method is more effective in complex real-world situations. Unlike many other supervised methods, Bayesian classification requires little training data to estimate model parameters.

2.8.3 Discriminative Approaches

Discriminative methods model the posterior probability $P(c_i|x)$ in a direct manner or learn a direct correspondence between the input data and the different target classes. There are several compelling reasons for favoring the use of discriminative methods to the detriment of generative methods, one of them is succinctly articulated by authors in (134), which say that we generally tend to solve the problem of classification directly and that we do not concern ourselves with a more general problem as an intermediate step. Regardless of computational issues and lack of data, the prevailing consensus states that discriminant methods are generally preferred over generative methods (135).

2.8.3.1 K-Nearest Neighbors

K-Nearest Neighbors (K-NN) is a well-known method in the learning and vision field (136). Unlike some approaches, the learning phase in K-NN is transparent. Indeed, K-NN requires memorization of all training data because it does not generalize. It consists of classifying an object by the majority vote of its neighbors. The most predominant class among its K-closest neighbors (K is typically little) will be appointed to it. If $K = 1$, the article is just doled out to the class of its closest neighbor; its name is improved for this situation to ”the closest neighbor” (1-NN).

From one viewpoint, this strategy has a few disadvantages, the time has come burning-through and memory-escalated in light of the fact that it requires stacking all of the preparation information, and working out the distances between each test model and all of the preparation models. This builds the calculation time. K-NN is likewise touchy to the presence of boisterous information, which makes it hard to sum up. One potential arrangement comprises in picking a subset of non-loud trademark vectors (137). This method comes up against another problem when certain classes are represented by few individuals. Indeed, the most frequent classes tend to dominate the rare classes in the prediction. Because of their large number, the ruling classes will be more represented in the K- nearest neighbors,

and therefore, they will affect the majority vote. To remedy this problem, one version of this method consists in weighting the vote of each of the K nearest neighbors by the distance separating it from the test example to be classified. It is prescribed to analyze the aftereffects of the new learning calculations with those of 1-NN on the grounds that the presentation of the last option approaches is steady and frequently great (138).

2.8.3.2 Kernel Methods

The kernel methods are especially requested in the case of non-linearly repairable data. This kind of approach is based on the "Mercer" theorem (139), which says that every kernel function is positive, semi-definite, symmetric, continuous, can be communicated as a dab item in a high-dimensional space.

The first application of kernels in machine learning dates back to 1964 with the work authors in (140), SVM "support vector machines" becomes the most illustrious kernel-based technique. To separate non-linearly separable data, these methods simulate the passage in a large-dimensional space. In this space, which it is not necessary to manipulate explicitly, linear methods can be implemented to find linear regularities there, corresponding to non-linear regularities in the original space. Thanks to the use of kernel functions, it becomes possible to have the best of both worlds: using simple and rigorously guaranteed techniques, and dealing with non-linear problems. This is why these methods have become very popular recently.

Support vector machines Backing Vector Machines (SVM) is one of the most well known strategies in the group of discriminative, piece based, order draws near. It was developed by authors in (141) in 1995, furthermore, stays right up 'til the present time perhaps the most utilized algorithm, particularly for design acknowledgment.

This is due to its ability for generalization and on the other hand to the notion of the kernel which makes it better suited to solve the problem of nonlinearly separable data. Thanks to these qualities, this method has been adopted by the image and video community, where the characteristic vectors of these data are large. The work in (142) is one of the latest exploration that applied SVM for picture order. Given two classes of D-dimensional information, the fundamental standard of SVM is to find an isolating hyperplane (s) impeccably recognizing the two information classes, by expanding the edge isolating the information of this hyperplane. Figure 2.7 outlines this rule on account of a two-dimensional space. H addresses a "isolating hyperplane" that different the dark circles from the white ones. As the information is for the most part not straightly detachable, SVM utilizes the idea of portion permitting to extend the information in a high dimensional space where the information will be directly distinct (143). Learning the boundaries of the separator hyperplane and the edges

is finished utilizing the improvement models. SVM expects to find the ideal hyperplane that augment the edge between the hyperplane and the help vectors. In view of the work in (144), RBF parts by and large give the best results.There are a few kinds of portion:

- **Linear (simple dot product)**

$$K(x_i, x_j) = x_i \dot{x}_j \tag{2.21}$$

- **RBF (Radial Basis Function)**

$$K(x_i, x_j) = exp(-\frac{\|x_i - x_j\|^2}{2 \times \sigma^2}) \tag{2.22}$$

- **Polynomial**

$$K(x_i, x_j) = (x_i \dot{x}_j + c)^2 \tag{2.23}$$

- **Sigmoïde:**

$$K(x_i, x_j) = tanh(x_i \dot{x}_j + c) \tag{2.24}$$

2.8.4 Ensemble Learning model

As a rule, there is nothing of the sort as a singular learning calculation that consistently prompts the most dependable model in any exploration field. Each learning calculation depends on a series of expectations, and on account of these presumptions it will not work with the thought about information, this leads to errors and very poor performance in terms of precision. To remedy this problem, some researchers have proposed methods that use not only one but a set of learning algorithms. These techniques are called: “Ensemble learning technique” (145). The idea of this type of method is to build a group of learners who, when combined, generate a “meta-model” that has better accuracy than individual learners. Basic learners are not selected for performance, but for simplicity. Several set-based learning methods have emerged, we can categorize them into four main families: “voting”, “bagging”, “boosting”, “stacking”. These methods differ in the way in which the decisions of the basic learners are combined.

2.8.4.1 Voting

This strategy consists of combining the decisions (predictions) of all individual learners:

$$y = f(d_1, d_2, \ldots, d_i, \ldots, d_n | \varnothing) \tag{2.25}$$

where:

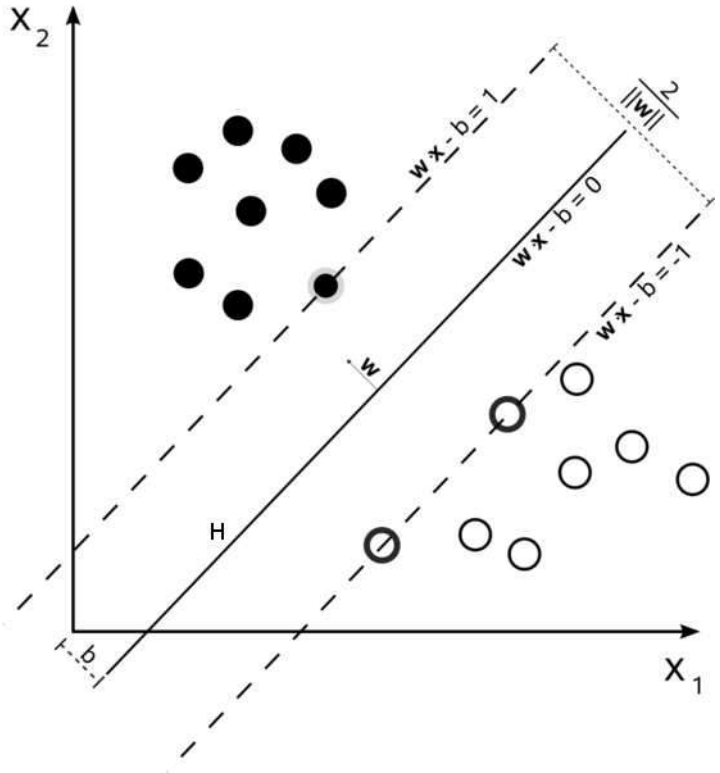


Figure 2.7: Straight division in two-dimensional space. with $\|.\|$ the L2 standard. x_i, x_j are two unmistakable vectors, and σ a Gaussian limit to be improved by cross-endorsement. This prompts a symmetric lattice called a "section structure", which shows the similarity between each pair of data vectors. Overall, just similarity limits which lead to a structure satisfying Mercer's conditions can be used

- n: is the number of learners considered.
- y: is the final prediction.
- $\varnothing$ is the set of parameters
- d_j is the decision of the j^{th} learner.
- f is the voting method.

The simplest voting technique is majority voting. The votes of individual learners can also be weighted by degrees of confidence or importance (weight). This method is called "weighted voting".

$$y = f(d_1, d_2, ..., d_i, ..., d_n|\varnothing) = \sum_{j=1}^{n} = d_j, w_j \tag{2.26}$$

with: $w_j > 0$ and $\sum_{j=1}^{n} w_j = 1$ where: w_j is the weight given to the j^{th} learner. The value of the weights can be related, for example, to the performance of the learners.

2.8.4.2 Bagging

Packing is the contraction for "Boostrap totaling". Packing was proposed by creators in (146) to further develop order execution by consolidating characterizations from an arbitrarily produced dataset. It is a "casting a ballot" type strategy in which students are prepared on somewhat various arrangements of information. For sure, for every student A, another learning dataset is created by arbitrarily drawing with bootstrap of the models from the first learning dataset. The choices of the piece students are then consolidated by a greater part vote. Any kind of order model can be utilized as a fundamental student. The sacking procedure is ordinarily utilized for supposed "shaky" learning calculations, where a little change in the learning information can cause a critical change in the subsequent model. Neural organizations and choice trees are great applicants that can profit from the upsides of Bagging. In (147), the creators proposed a SVM-based Bagging plan with a one-sided determination of positive and negative examples to resolve the issue of uneven classes with regards to ordering interactive media records.

multi-SVMs: the authors in (147) proposed a method which consists of combining m classifiers via a "bagging" strategy where each of them uses all the learning samples of the dominated class (typically, the positive class) and a set of dominant class samples (typically the negative class) are drawn randomly with bootstrap, with:

$$m = \frac{f_{neg} \times N_{neg}}{f_{pos} \times N_{pos}} \tag{2.27}$$

where N_{pos} is the number of positive examples, $Nneg$ is the number of negative samples, $fneg$ and f_{pos} are parameters (non-zero positive integers) related to the positive and negative classes, respectively. We may say at this stage that the annotation concerns a pair of concepts and not an individual concept. f_{pos} manages the proportion of the samples of the dominant class that we want to use, compared to the number of samples of the dominated class. f_{neg} with the use f_{pos} allows to control the number of the desired classifiers. The set ED is divided into m subsets, where every subset contains every one of the positive examples contained in ED and $(f_{pos} \times N_{pos})$ negative examples are drawn haphazardly with rebate. Then, at that point, each of the m classifiers is prepared on an alternate subset. Note that the limitation $f_{neg} \times N_{neg} - f_{pos} \times N_{pos}$ should be confirmed. At long last, the scores of the m classifiers are consolidated utilizing any conceivable capacity, regularly a normal. The more prominent the worth of m, the better the last execution.

2.8.4.3 Boosting

Boosting technique consists of improving learners having a weak performance. It has been proven in (148) that it is possible to transform such learners into good learners who can classify un-annotated samples in a good manner. The similarity between bagging and boosting only boils down to building a set of classifiers by sampling the training dataset, and combining the decisions of different classifiers by majority vote.

The sampling of learning subsets is done in boosting so as to provide for the next learner as much as possible the most informative data set. Among the most famous boosting methods, we can cite "Adaboost" proposed by authors in (149). This algorithm gives weight to all of the learning samples. At every emphasis i, a classifier C_i is prepared to limit the order mistake. This blunder is determined and utilized by C_i to refresh the weight circulation of the preparation test. This update of the weights of samples is done in such a way that the bad-classed samples in the current iteration have a better chance of being introduced into the training set of the classifier of the next iteration. The process is iterated until a stop criterion, usually related to the error rate, is verified. Several generalizations of Ada-Boost to the multi-class case have emerged (150).

2.8.4.4 Stacking

Stacking is a group learning technique that is very similar to the "voting" method, it was proposed by authors in (151). Stacking consists of combining several classifiers. Initial, various classifiers are prepared utilizing the preparation dataset, their outputs are then combined using a new meta-classifier that matches the decisions of the basic classifiers individuals to those correct classes of samples. In multimedia indexing, the stacking technique is widely used as a late merging method. It is an efficient strategy that to merge the scores of classifiers, which are obtained from different modalities (152).

2.8.5 Background Neural Networks

As of late, profiting from the fast improvement of profound learning innovation, the computer vision field has achieved unprecedented success. There are many algorithms and techniques developed in that field. In the following, we will detail each learning method.

2.8.5.1 Convolutional Neural Networks

As quite possibly the most broadly utilized neural network, Convolutional neural organizations (CNNs) are the center learning calculations for visual example acknowledgment. They were created from perceptrons, vector planning calculations enlivened by affiliated learning

of the mind, and the possibility of “integrate and fire” neurons. Researchers have employed CNNs and their variants (e.g., ResNet) to tackle a variety of challenges such as image classification, object recognition, action recognition, pose estimation, neural style transfer, etc. Previous studies have shown that they outperform humans in some recognition tasks.

CNNs are made out of various neural units, which can be for the most part separated into three sorts, in particular, the info layer, the secret layer, and the yield layer. The information layer of a convolutional neural organization is predominantly used to acquire input data, which can process multi-dimensional information.

The yield layer of a convolutional neural organization typically utilizes a consistent capacity or softmax capacity to yield the order names. The useful application fluctuates as per the sort of assignment. For example, the yield layer can be the focal directions, size, and arrangement of items in object recognition. In semantic picture division, the yield layer straightforwardly yields the arrangement name of every pixel.

In general, each neural unit in the input layer directly connects to the original data and provides feature information to the hidden layer. Each neural unit in the hidden layer represents different weights for different neural units in the input layer, so it tends to be sensitive to a certain recognition pattern. The values in the output layer vary according to the activation degree of hidden layers, which is the final recognition result of the model. Contrasted and the information layer and yield layer, the secret layer is more mind boggling on the grounds that it is intended for conceptual component extraction. It generally incorporates the convolutional layer, pooling layer, and completely associated layer.

2.8.5.2 Artificial Neural Networks (ANNs)

ANNs are figuring frameworks intended to imitate the human mind’s data handling component. Such frameworks ”learn” to execute undertakings by considering models without being modified with any assignment rules and they have self-learning abilities that make them produce better outcomes when more information become accessible. For instance, assuming somebody needs to recognize pictures that contain salad dishes, he can utilize model pictures that have marked as ”salad” or ”no plate of mixed greens” to prepare an ANN and utilize the prepared organization to distinguish salad in other new pictures. ANNs play out the errands with next to no earlier information about salad dishes. All things considered, they naturally produce distinguishing attributes from the learning material (e.g., named salad pictures).

An ANN comprises of an assortment of associated hubs called counterfeit neurons, which model the neurons in a human mind. Every association can communicate a sign starting with one fake neuron then onto the next. The fake neuron (signal recipient) can handle it and afterward communicate it to the associated fake neurons. In like manner ANN executions,

the sign at an association between fake neurons is a genuine number, and the yield of each fake neuron is registered by some nonlinear capacity of the amount of its bits of feedbacks. The associations between counterfeit neurons are called 'edges'. Counterfeit neurons and edges ordinarily have a weight that changes as learning continues. The weight increments or diminishes the strength of the sign at an association. Counterfeit neurons might have a limit with the end goal that the sign is possibly conveyed assuming the total message passes that boundary.

2.8.6 Neural Network Layers

Regularly, fake neurons are totaled into layers. Various layers might perform various types of changes on their bits of feedbacks. Signs travel from the main layer (the info layer) to the last layer (the yield layer), conceivably subsequent to intersection various layers.

2.8.6.1 Convolution Layer

Convolutional layers are the essential structure blocks utilized in convolutional neural organizations. The convolution as a channel empowers the neural organization to extricate compelling undeniable level highlights. The element map, additionally called the initiation map, can be created by over and over applying a similar channel, which demonstrates the areas and strength of recognized highlights in the information picture. The channel contains the loads that should be gotten the hang of during the preparation of the layer. Additionally, the channel size or bit size will essentially influence the state of the yield highlight map. It is important that the connection of the channel with the line of the picture might prompt boundary impacts, particularly for the little size input picture and exceptionally profound organization. Us-partner, we can fix the boundary impact issue by adding additional pixels to the edge of the picture, which is called cushioning. Moreover, the measure of development between the channel applications to the info picture is alluded to as the step, and it is quite often balanced in stature and width aspects. For instance, the step (2; 2) implies moving the channel two pixels appropriate for every even development of the channel and two pixels down for every upward development of the channel while making the element map. The step of the channel on the information picture can be viewed as the down-examining of the yield include map. Then, we present three fundamental properties of the convolutional layers: meager connections, boundary sharing, and equivariant portrayals.

- **Sparse interactions**: Convolutional neural organizations have inadequate cooperations by making the portion more modest than the information. When handling a picture with huge number of pixels, we can distinguish little important elements like

the edges of the picture by taking just tens to many pixels. This not just lessens the capacity prerequisites of the model yet additionally further develops its general calculation effectiveness.

- **Parameter sharing**: Parameter sharing is the sharing of loads by all neurons in a specific component map. Every neuron is associated uniquely to a subset of the information picture, which is additionally called neighborhood network. This property assists with decreasing the quantity of boundaries in the entire framework and makes the calculation more effective.

- **Equivariant representations**: For convolution activity, boundary sharing makes the neural organization layers have an equivariant portrayal. In particular, assuming the info is somewhat moved, the aftereffect of the convolution activity is something very similar. Note that the convolution isn't normally identical to some different changes, for example, picture scaling or pivot change.

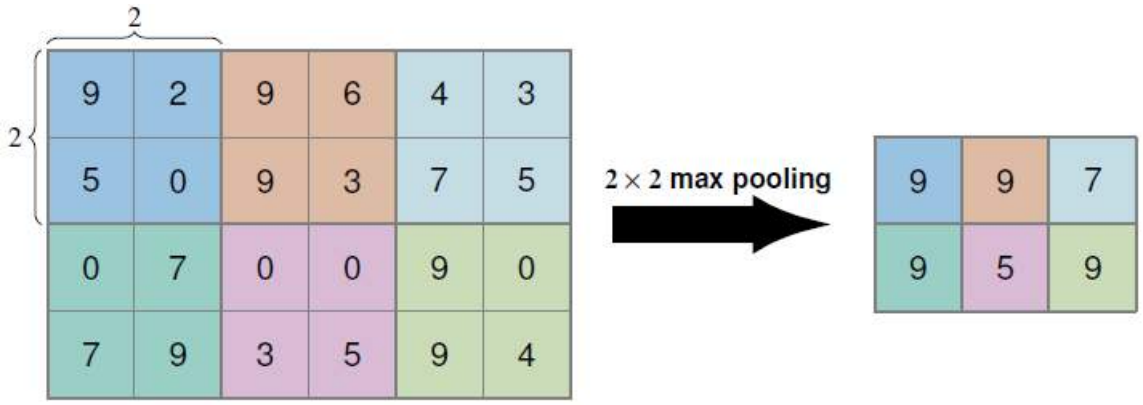


Figure 2.8: Example of max pooling operation

2.8.6.2 Pooling Layer

Pooling operation plays a vital role in the structure of convolutional neural networks. First of all, pooling layers improve the spatial invariance to some extent, such as translation invariance, scale invariance, and deformation invariance. Namely, even if the image input is transformed slightly, the pooling layer can still produce similar pooling features, making the learning system more robust. Secondly, pooling operation is equivalent to feature downsampling, which increases the receptive field size. For some visual tasks, a large receptive field helps learn long-range spatial relationships and implicit spatial models. In addition,

pooling operation greatly reduces the model parameters, which leads to a lower risk of overfitting. Assume that the element of the picture input is $c \times w \times h$, where c is the quantity of channels, and w and h are the width and stature, individually. Assuming the step of the pooling layer is set to 2, the component of the yield picture will be $c \times w = 2 \times h = 2$. For this situation, both the computational expense and memory utilization will be decreased by a component of 4.

Common pooling methods include **average pooling** and **maximum pooling**. Maximum pooling calculates the maximum value of the target patch, which retains more texture information of the image input, whereas average pooling keeps more background information and tends to transfer the comprehensive information in the architecture of convolutional neural networks. Figure 2.8 shows an example of 2×2 maximum pooling.

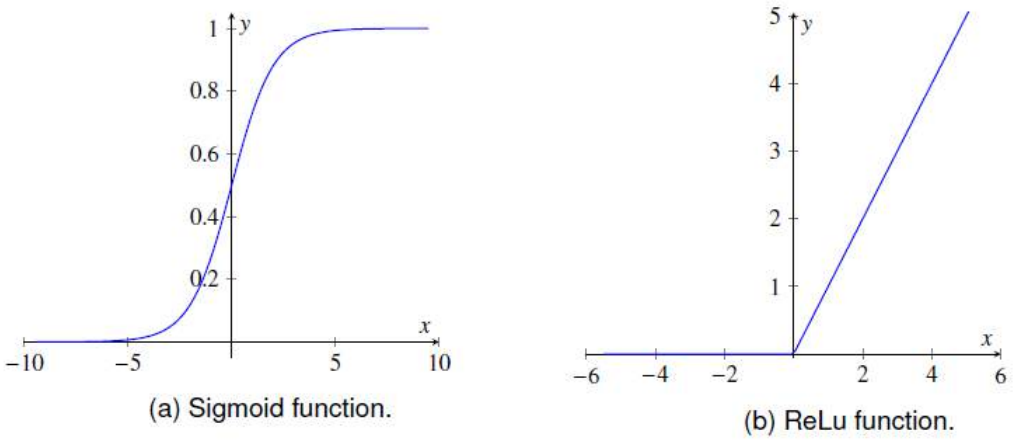


Figure 2.9: Examples of popular activation functions

2.8.6.3 ReLU Layer

In neural organizations, the enactment work plays out the nonlinear change to the info, making it fit for learning and performing more intricate errands. To build the nonlinearity of neural organizations, some nonlinear capacities are presented. Clearly, the amassing of various straight capacities is as yet direct, while straight capacities have restricted articulation. The utilization of nonlinear capacities makes the organization more expressive and along these lines better fits the objective capacity. Two normal nonlinear capacity utilized in convolutional neural organizations are the **sigmoid function** and the **rectified straight unit** (ReLU).

As shown in Figure 2.8, we can observe that the ReLU activation function has more advantages than the sigmoid function. ReLU can carry out negative suppression so as to be more sparsely active. More importantly, ReLU activation functions suffer less from the

vanishing gradient problem. The derivative of the sigmoid function has good activation only when it is near zero. The gradient in the positive and negative saturation region is close to zero. Also, the derivative of the ReLU function is easy to calculate, which can accelerate the model training to some extent.

2.8.6.4 Fully Connected Layer

The fully connected layer is usually used as a classifier to connect the hidden layer and the final output. In the architecture of convolutional neural networks, adding several fully connected layers after the convolution layers can map the generated feature map into a fixed-length feature vector. The final output represents the numerical description of the input image. This structural property is conducive to the realization of image-level classification and regression tasks.

Although multiple fully connected layers can significantly improve the nonlinear expression ability of learning models, a large number of neurons increase the model complexity. Plenty of model parameters will reduce the efficiency of the learning algorithm and even lead to overfitting. Therefore, the trade-off between accuracy and efficiency has been deeply explored in deep learning technology research. For the segmentation task, however, spatial information should be stored to make a pixel-wise classification. Hence, the fully connected layer is usually substituted by another convolution layer with a large receptive field.

2.8.7 Optimization

Improvement calculations are significant for profound learning. On one hand, preparing an intricate profound learning model can require hours, days, or even weeks. The presentation of the streamlining calculation straightforwardly influences the model's preparation effectiveness. Then again, understanding the standards of various advancement calculations and the job of their hyperparameters will empower us to tune the hyper-boundaries in a designated way to work on the exhibition of profound learning models. There are many profound learning streamlining calculations in the writing, we decide to discuss the most known ones; clump Normalization and dropout since we expect to utilize them in our execution.

2.8.7.1 Batch Normalization

Preparing profound neural organizations with various secret layers is very difficult. One explanation is that the model is refreshed layer-by-layer in reverse from the yield to the information utilizing a mistake gauge that accepts the loads in the layers before the current layer are fixed. This dials back the preparation by requiring lower learning rates and cautious boundary introduction and makes it famously difficult to prepare models with immersing

nonlinearities (153). Consequently bunch standardization, as a powerful enhancement strategy, is proposed to normalize the contributions to a layer for every smaller than expected group while preparing extremely profound neural organizations. It balances out the learning system and significantly lessens the quantity of preparing ages needed to prepare profound organizations. Bunch standardization can be carried out during preparing by working out the mean and standard deviation of each information variable to a layer for every scaled down clump and utilizing these measurements to play out the normalization. On the other hand, a running normal of mean and standard deviation can be kept up with across small scale clumps however may bring about unsteady preparing. For Example, creators in (153) utilized clump standardization after the convolutional layers in their exceptionally profound model, alluded to as ResNet. The revealed results accomplished cutting edge in the picture characterization task. In our work of model plan, we ordinarily add group standardization change before nonlinearity.

2.8.7.2 Dropout

Profound neural organizations are probably going to get overfitting while at the same time preparing with not many models. As ahead of schedule as 2012, creators in (154) have proposed the idea of dropout, which is currently broadly utilized in cutting edge neural organizations. Probabilistically exiting hubs in the organization is a basic and compelling regularization technique (155). In every cycle, a few hubs are arbitrarily erased, and just the excess hubs are prepared. This enhancement technique diminishes the connection among's hubs and the intricacy of the model to accomplish the impact of regularization. By and large, dropout just necessities to set a promotion boundary that is the extent of hubs haphazardly safeguarded in each layer. In particular, the boundary grid of this layer is determined with the double network created by the hyperparameter by means of the point-by-point item.

2.8.8 Model Training

The process of training neural networks is the most challenging part of using deep learning techniques and is by far the most time-consuming, both in terms of effort required for configuration and computational complexity required for execution. In the following, we summarize the commonly used techniques in model training, including data preprocessing, weight initialization, loss function, and gradient descent optimization.

2.8.8.1 Data Preprocessing

For the most part, preparing profound learning models requires a great deal of information due to the immense number of boundaries should have been tuned by the learning calcu-

lation. Information preprocessing is an essential assurance for viably model preparing, and we should be mindful so as to set up the preparation information to accomplish the best expectation results. For instance, many profound learning models have standardized info handling, in particular the brightening activity, which changes the normal pixel worth of the picture to nothing and the difference of the picture to unit fluctuation. Exhaustively, the mean and fluctuation of the first picture are first determined, then, at that point, every pixel worth of the first picture is changed. This activity empowers the combination of the neural organization quicker. Normal information preprocessing techniques additionally incorporate information quality appraisal, highlight collection, include examining, dimensionality decrease, and element encoding.

Furthermore, information expansion (156) is much of the time utilized in model preparing, which builds the measure of information by adding somewhat adjusted duplicates of previously existing information or recently made engineered information from existing information. In genuine situations, we might have a limited scale dataset of pictures taken in a restricted arrangement of conditions. On account of restricted information, information expansion can build the variety of preparing tests, in order to work on the vigor of the model and stay away from overfitting. Average tasks incorporate flipping, pivot, shift, resize, arbitrary scale, irregular harvest, shading jittering, contrast, clamor, extravagant PCA, GAN, and so forth

2.8.8.2 Weight Initialization

Training a deep learning model means learning good values for all the weights and the bias from labeled examples. In particular, the bias allows to shift of the activation function by adding a constant. In order to consistently update the weights, the models require each parameter to have the corresponding initial value. For convolutional neural networks, the nonlinear function is superimposed by multiple layers, and how to select the initial value of parameters becomes a problem worthy of discussion.

By and large, the motivation behind weight instatement (157) is to forestall the layer initiation yield from detonating or vanishing in the forward move cycle of profound neural organizations. Regardless, the misfortune angle is either excessively enormous or too little to even consider streaming in reverse beneficially. The learning model will then, at that point, set aside a more drawn out effort to unite. Additionally, it is striking that instating every one of the loads with zeros drives the neurons to gain proficiency with similar highlights during preparing. The model can not get the update of boundaries accurately. For instance, expect we introduce every one of the inclinations to nothing and the loads with some steady β. On the off chance that we forward spread an information $(x_1; x_2)$ in the organization, the yield

of stowed away layers will be relu($\beta x_1 + \beta x_2$). In particular, the secret layers will impact the expense, which will bring about indistinguishable inclinations.

Practically speaking, scientists generally utilize the Xavier introduction (157) to keep the difference the equivalent across each layer. Another normal instatement is He introduction (158) in which the loads are instated by increasing by two the difference of the Xavier instatement.

2.8.8.3 Loss Function

The neural organizations are typically prepared utilizing the inclination plunge improvement calculation. By and large, the streamlining issue includes a true capacity that demonstrates the course of improvement. During the improvement interaction, the organization attempts to track down an up-and-comer answer for boost or limit the goal work. Under requirement conditions, we ascertain and limit the model blunder by means of a misfortune capacity or cost work, which assesses the attack of the learning model. This series of limitations is a regularization term that assists with forestalling overfitting. The loads are refreshed utilizing the back-engendering of the blunder calculation. Along these lines we want to pick an appropriate misfortune work when planning and arranging the model. Assume that there is a progression of preparing tests $(x_i, y_i)_{i=1,\dots,N}$ in regulated learning.

The model learns the planning connection of $x \to y$, so that given a x, regardless of whether the x isn't in the preparation tests, it can get the yield $\hat{y}$ as near the genuine y as could really be expected. The misfortune work is a vital part to demonstrate the bearing of model improvement, which is utilized to gauge the contrast between the yield $\hat{y}$ of the model and the genuine yield $\hat{y}$, to be specific, $L = f(y_i; \hat{y}_i)$.

In the accompanying, we present a few misfortune works generally utilized for characterization and relapse in profound picking up, including mean squared blunder misfortune, mean outright mistake misfortune, and cross-entropy misfortune. By and by, analysts ordinarily utilize the Xavier introduction (157) to keep the fluctuation the equivalent across each layer. Another normal introduction is He instatement (158) in which the loads are instated by increasing by two the fluctuation of the Xavier instatement.

1. **Mean Squared Error Loss**: Mean Squared Error Loss (MSE), otherwise called Quadratic Loss or L2 Loss, is the most normally utilized misfortune work in AI and profound learning relapse undertakings. Officially,

$$J_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2.28}$$

Under the suspicion that the blunder between the yield of the model and the genuine worth follows a Gaussian appropriation, the base mean square mistake misfortune work and the greatest probability gauge are basically reliable. Accordingly, in the situation where the suspicion can be fulfilled (like relapse), the mean square blunder misfortune is a decent decision for the misfortune work. In situations where this supposition that isn't fulfilled (like order), different misfortunes must be thought of.

2. **Mean Absolute Error Loss**: Mean Absolute Error Loss (MAE), otherwise called L1 Loss, is another normal misfortune work. MSE misfortune by and large meets quicker than MAE misfortune, notwithstanding, the last option is more strong to the exception, which can be characterized as

$$J_{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{2.29}$$

3. **Cross-Entropy Loss**: Previous misfortune capacities acquainted above are relevant with relapse issues. For order errands, particularly for semantic picture division, the most usually utilized misfortune work is the cross-entropy misfortune that increments as the anticipated likelihood wander from the genuine mark. For the multi-class cross-entropy misfortune, we can get:

$$J_{CE} = -\sum_{i=1}^{N} y_i^{c_i} log(\hat{y}_i^{c_i}) \tag{2.30}$$

Here, $y_i^{c_i}$ can be 0 or 1, demonstrating whether class mark c_i is the right order.

2.8.9 Deep Learning

Deep learning was originally proposed by G. E. Hinton in 2006 (159) for the representation of data (image, audio, text, etc.) by mimicking the multilayered abstraction mechanism of the human brain. They wrote a system that combines feature learning and classification into a single learning process. CNNs can be used for unsupervised learning of features and can therefore be used as tools to generate low-level descriptors. Indeed, with their hierarchical multilayer architecture, CNNs are able to learn and recognize visual patterns directly from the pixels of images. They can also be used as learners outputting the classes of the sample while doing combined learning of the characteristics and the separation between the different classes. Moreover, CNNs are well known for their adaptation to minimal preprocessing and their robustness to (160) distortion.

In this Section, we present profound learning since we utilized the profound model in the article characterization. Profound learning is a subset of AI techniques dependent on

counterfeit neural organizations (161); it utilizes different layers to separate more significant level elements from the crude info (162). Profound learning models, for example, convolutional neural organizations (CNN)s and repetitive neural organizations (RNN)s have been applied to fields including regular language handling, PC vision, discourse acknowledgment, and sound acknowledgment, where they have created results similar to human specialists (22, 163).

In profound learning, each layer figures out how to change its feedback information into a more dynamic portrayal. For instance, in face acknowledgment application, the crude info is a lattice of pixels; the main illustrative layer will digest the pixels and encode edges; the subsequent layer might encode plans of edges; the third layer may en-code a nose and eyes, and the fourth layer might perceive that the picture contains a face (164).

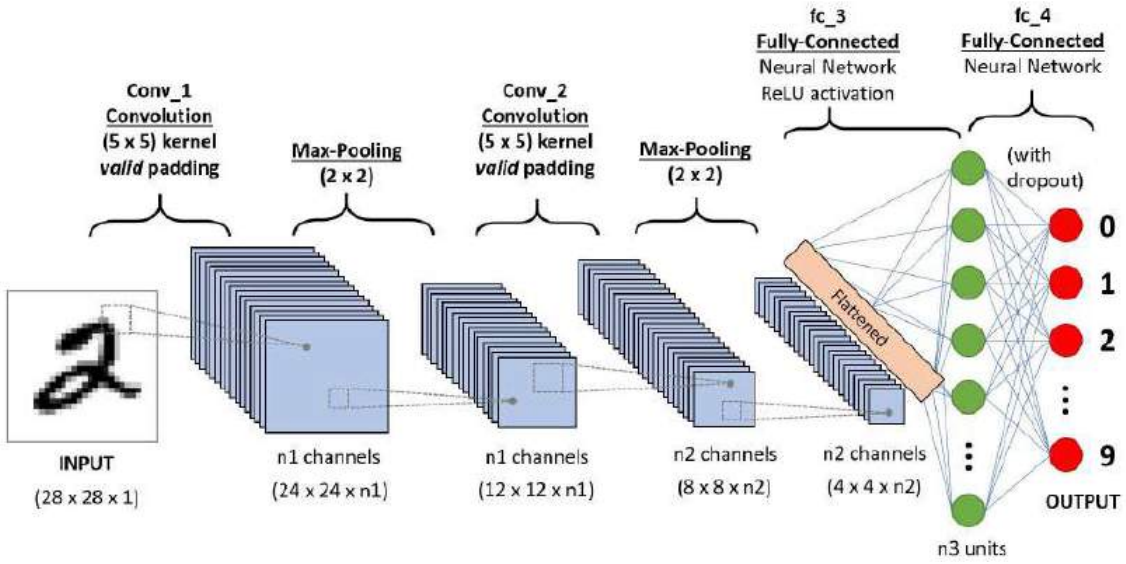


Figure 2.10: A CNN sequence to classify handwritten digits (165)

2.8.10 Transfer Learning

Many AI techniques function admirably under a typical presumption: the preparation and test information are drawn from a similar component space and a similar appropriation. At the point when the appropriation changes, most factual models should be modified without any preparation utilizing recently gathered preparing information (166). In some genuine applications, it is costly or difficult to re-gather the required preparing information and remake the models. It would be great to decrease the need and work to remember the preparation information. In such cases, information move or move learning between task spaces would be attractive.

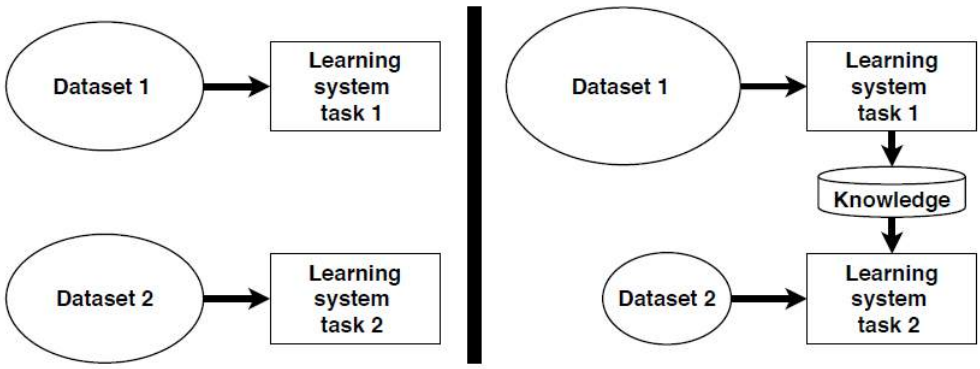


Figure 2.11: On the left learning collaboration of regular AI; On the right learning cycle of move learning

Move learning is an AI strategy where a model created for an errand is reused as the beginning stage for a model on a subsequent assignment as displayed in Figure 2.11. It is a well known methodology in profound realizing where pre-prepared models are utilized as the beginning stage on PC vision and regular language handling undertakings given the immense process and time assets needed to foster neural organization models on these issues and from the gigantic leaps in an expertise that they give on related issues (167).

As displayed in Figure 2.11, conventional learning is secluded and happens simply dependent on explicit errands, datasets, and preparing separate disconnected models on them. No information is held which can be moved starting with one model then onto the next. In move learning, you can use information (highlights, loads, and so on) from recently prepared models for preparing more up to date models and even tackle issues like having less information for the fresher assignment.

2.8.10.1 Transfer Learning Strategies

In move learning, we want to figure out what piece of information can be moved across spaces or undertakings. In the wake of finding which information can be moved, we want to foster learning calculations to move the information. In light of various circumstances between the source and target assignments and spaces we can classify the exchange learning into inductive exchange learning, transductive exchange learning and, unaided exchange learning.

In the inductive exchange getting the hang of setting, the objective errand is not quite the same as the source task however they are connected, regardless if the source and target areas are something similar or not. In the transductive exchange getting the hang of setting, the source and target errands are something similar, while the source and target spaces are unique. At long last, for the solo exchange getting the hang of setting the objective assignment is not quite the same as the source task however they are connected, like the inductive exchange

picking up setting. Nonetheless, the unaided exchange learning center around tackling solo learning undertakings in the objective space, for example, grouping (164), dimensionality decrease and thickness assessment (168).

2.8.10.2 Deep Transfer Learning Srategies

Profound learning has gained momentous headway as of late. This advancement has empowered analysts to attempt confounded issues and yields astounding outcomes. Nonetheless, the necessary measure of information and the preparation time for such profound learning frameworks are significantly more than contrasting and the conventional ML frameworks. There are different profound learning networks with the best in class execution that have been created and tried across fields like PC vision and regular language handling. By and large, individuals share the subtleties of these organizations for others to utilize. These pre-prepared models structure the premise of inductive exchange learning with regards to profound learning (profound exchange learning). The two most regularly utilized profound exchange learning techniques are:

- **Train the whole model**—utilize the executed design of the pre-prepared model and train it on your dataset. Rather than utilizing arbitrary loads, start from upsides of a pre-prepared model.
- **Feature extraction (freezing CNN model base)**—train another classifier on top of the pre-prepared base model. The loads of convolution layers are left unaltered and just the last, completely associated layer is prepared.
- **Fine-tuning (preparing likewise some convolution layers)**—retrain at least one convolution layers notwithstanding a completely associated classifier. Unique convolution layer loads are utilized as beginning stages. Opened convolution layers are simply tuned to another issue.

The key thought is to utilize the pre-prepared model's weighted layers as highlights extractor without refreshing the loads of the model's layers during the new assignment's preparation stage.

Profound learning models' exhibition expands relatively to the measure of information (169). Along these lines, most designs must be prepared on extremely huge datasets, which, regularly, are not promptly accessible. Notwithstanding, the most famous models are given on an open-source premise, with pre-prepared loads. They are prepared on the ImageNet dataset (170) an assortment of 14,197,122 pictures from 21,841 genuine classes. This dataset is utilized in overall rivalries, from which, consistently, better calculations arise.

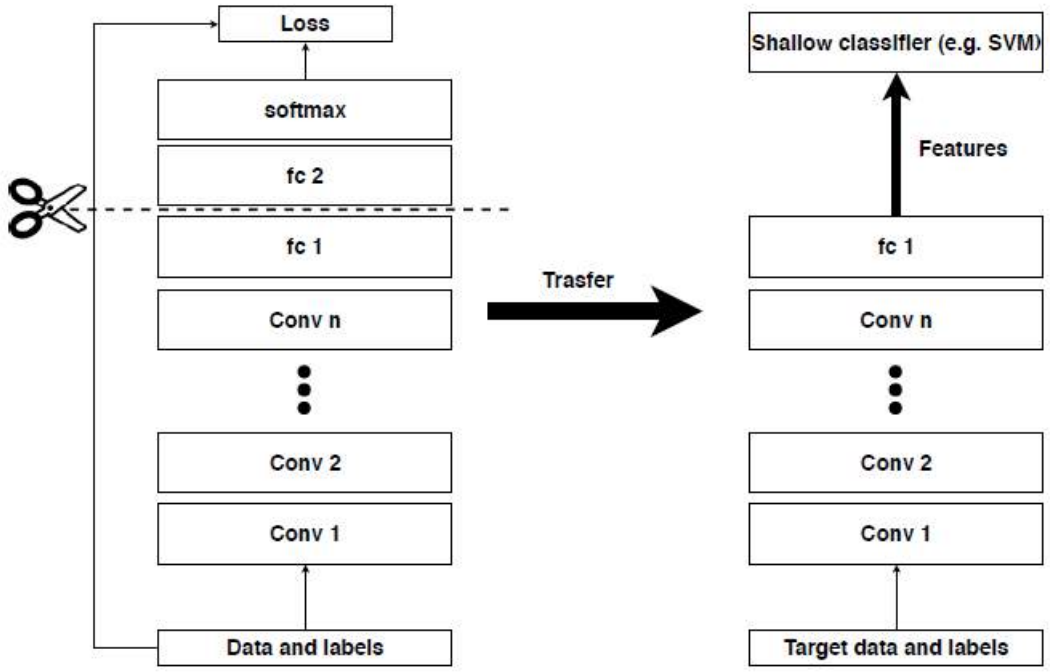


Figure 2.12: Move Learning with Pretrained Deep Learning Models as Feature Extractors

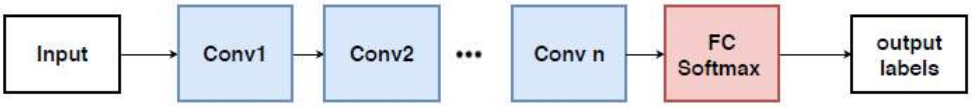


Figure 2.13: In fine-tuning process, all convolutional layers (blue layers) in the organization are fixed and slope is backpropagated through the completely associated (FC) layer as it were

2.8.11 Prertained Deep Learning Architectures

Lately, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) has been overwhelmed by cutting edge CNNs and profound learning strategies, including pre-prepared organizations. Beginning from the popular AlexNet in 2012 (22), further developed designs, for example, VGG-19, ResNet50, Xception, and DenseNet121 addressed the absolute most noteworthy performing procedures in the course of recent years (171). These models, close by pre-prepared loads and helpful capacities, are given by the Keras library [1].

- **GG16 and VGG19** :(172)—probably the least difficult engineering comprising of just 3×3 convolutional layers stacked on top of one another. Lessening volume size is dealt with by max pooling (Figure 2.15).

- **ResNet**(153)—not at all like customary consecutive organizations like VGG, ResNet presents an organization in-network engineering. The ResNet model is worked from

[1] https://github.com/fchollet/keras

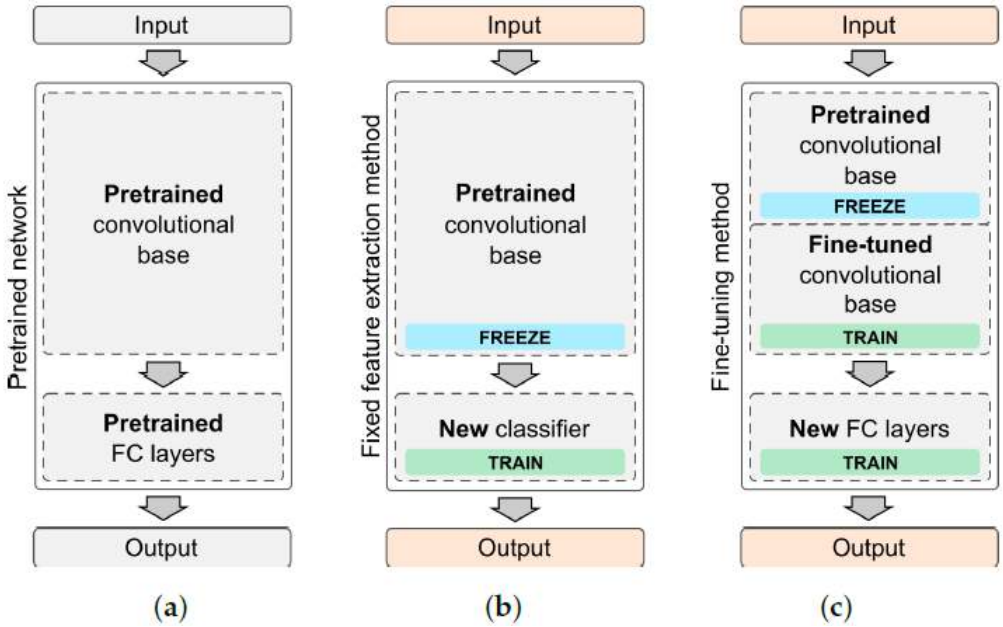


Figure 2.14: There are three transfer learning situations: (a) train entire model, (b) apply over again classifier on top of a pretrained convolutional base, (c) fine-tune the base, by re-preparing oneor more convolution layers

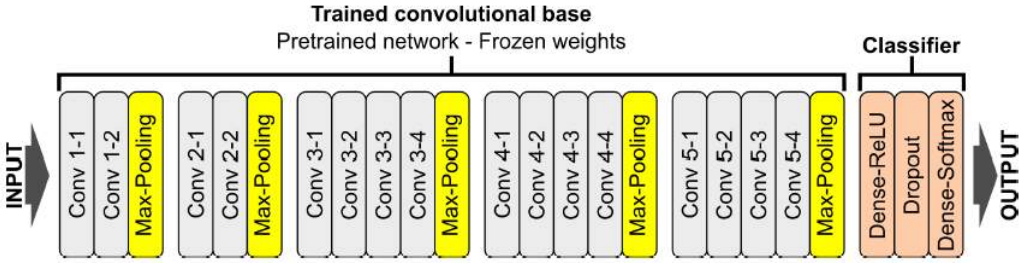


Figure 2.15: Schematic outline of the customized VGG-19 organization design with depiction of layers

layers stacked on top of each other and alternate ways at each layer that associate the contribution of that square with the yield. On account of the proposed arrangement, comprising of an-Identity Blocks and b-Convolutional Blocks, it was feasible to carry out an extremely profound neural organization with skipping-associations that assisted with tending to the evaporating angle issue (Figure 2.16a).

- **Inception** (175)—in this model, layers are regularly associated in equal as opposed to being stacked on top of one on another (Figure 2.16b). Complex channels are isolated into different basic ones. One of the changes, called Xception (176), utilizes profundity savvy distinguishable convolutions and 1×1 pointwise convolutions to diminish highlight

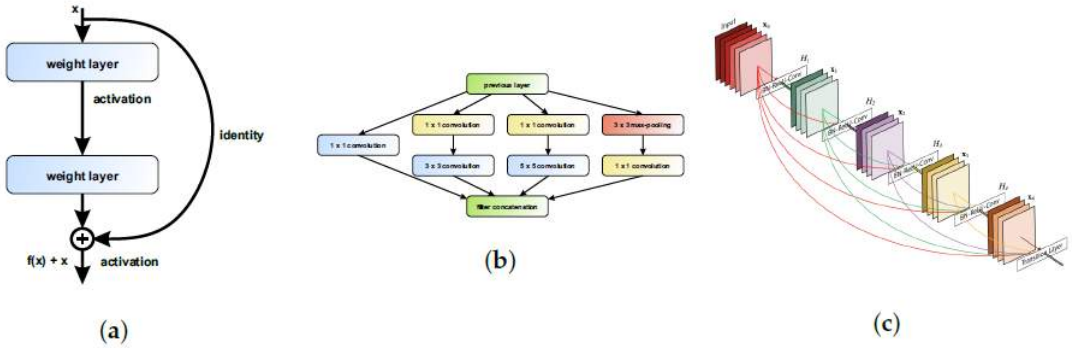


Figure 2.16: CNN-based neural organization engineering component: (a) an example remaining square clarifying the thought behind ResNet model (153); (b) the overall thought of Xception network (173); (c) schematic chart of DenseNet design (174)

Model	Dropout	Activation Function	Optimizer	Epochs	Batch
VGG-19	0.7	Sigmoid	AdaMax	30	128
ResNet-50	0.5	ReLU	AdaMax	20	512
Xception	0.5	ReLU	AdaMax	20	512
DenseNet121	0.5	Sigmoid	AdaMax	30	256

Table 2.1: Hyperparameters and preparing setings for the CNN models.

map aspects [1] (177).

- **DenseNet** (174)—the most recent organization, where each layer is associated with all past ones. Elements are determined based on data removed at every single past stage (Figure 2.16c). One of the essential parts of CNN structures are initiation works that are liable for the yield of a solitary neuron, and furthermore the exactness and computational productivity of the preparation model [2] (178).

2.8.12 Parameter Selection

The fundamental condition for legitimate neural organization preparing is the right determination of hyperparameters. To accomplish this, we performed matrix inquiry enhancement, with 5-crease cross-approval on the preparation dataset. We tried ReLU, sigmoid, TanH, and straight actuation capacities for completely (thickly) associated layers in the classifier. The dropout rate, which decides the small portion of the information units to drop, shifted somewhere in the range of 0.3 and 0.7. The boundaries that gave the most encouraging outcomes are introduced in Table 2.1.

[1] https:/openreview.net/pdf?id=S1jBcueAb
[2] https://towardsdatascience.com/activation-functions-in-neural-networks-eb8c1ba565f8

Preparing hyperparameters included: the analyzer type (SGD, RMSprop, Adagrad, Adadelta, Adam, AdaMax, Nadam), the clump size (32, 64, 128, 256, 512), and the quantity of ages (5, 10, 15, 20, 30). In the table, we present the boundaries that accomplished the most elevated precision (Table 2.1). As each streamlining agent works best with an alternate learning rate, we chose to adhere to default esteems set in Keras [1].

In our execution, the AdaMax analyzer, which is an adjustment of a famous streamlining calculation called Adam, gave the most noteworthy outcomes. As expressed in (179), the AdaMax calculation figures individual versatile taking in rates for various boundaries from evaluations of the first and second snapshots of the angles.

By actually looking at the exactness and loss of preparing and approval sets, we had the option to control the model's exhibition during preparing. Figure 2.15 shows that the ResNet50 and Xception models overfit the information for each situation, their preparation exactness increments, while the approval precision continues as before. Then again, DenseNet engineering gradually improves with every age, prompting a superior end-product. Eventually, we prepared each of the four organization models multiple times, with ideal boundaries. The eventual outcomes were confirmed on a different test set and found the middle value of.

2.8.13 Image Augmentation for Deep Learning

Picture increase is one more answer for beat the predetermined number of accessible commented on pictures. It builds the size of the accessible information by applying some picture change tasks to the current pictures from a preparation dataset to deliver new forms of existing pictures. Picture change tasks incorporate revolution, shearing, interpretation, zooming, and so on these irregular changes will create various pictures each time. (156)

2.8.14 Evaluation Metrics

Model assessment is a significant part when creating information driven models. The motivation behind any prescient model is to effectively foresee the objective class an incentive for concealed information occurrences with the most elevated conceivable precision. Consequently, it is needed to have a method of assessing model execution, commonly by evaluating it utilizing some proportion of model blunder. This equivalent measure should be utilized to prepare the model to get high exactness execution. One of the huge traps while making a prescient model is assessing the prepared model on something very similar or practically comparative information to the preparation ones. Taking on mistaken measures and assessment techniques might prompt create overfitted and over-hopeful models.

[1] (https://medium.com/octavian-computer based intelligence/which-analyzer and-learning-rate-should-I-use-fordeep-learning-5acb418f9b2)

2.8.14.1 Segmentation Evaluation

The division execution is by and large impacted by many elements, like the preprocessing of information, combination procedure, the decision of the spine organization, the act of best in class profound learning advances, and so on Hence how to assess and think about the presentation of division calculations is a basic issue. The legitimacy and handiness of a learning framework can be estimated in numerous angles, for example, execution time, memory impression, and exactness. It is prominent that current enormous scope benchmark datasets advance the normalization of examination measurements, giving a reasonable correlation f the best in class strategies. As a general rule, precision is the most well-known assessment standards to gauge the presentation of pixel-level expectation (180). For multimodal picture division, the most well known measurements are not quite the same as those utilized in unimodal methodologies, including Pixel Accuracy (PA), Mean Accuracy (MA), Mean Intersection over Union (MIoU), and Frequency Weighted Intersection over Union (FWIoU), which are first utilized in (181).

The IoU is characterized as $IoU = TP = (TP + FP + FN)$, where TP, FP and FN mean genuine up-sides, bogus up-sides and bogus negatives, separately. TP is the place where the underlying and the anticipated ones are both have a place with the genuine class. FN is the place where introductory is valid and predicated is bogus. TN is the place where the underlying and the predicated are both have a place with bogus class. FP is the place where the underlying is starting is bogus and the predicated is valid. For clarification, we signify n_{ij} as the quantity of pixels having a place with class i which are ordered into class j, and we consider that there are n_{cl} classes, $t_i = \sum_j n_{ij}$ is the quantities of pixel in class i. Consequently we can characterize these precision measurements as follows:

- Pixel Accuracy

$$\frac{\sum_i n_{ii}}{\sum_i t_i} \tag{2.31}$$

- Mean Accuracy

$$\frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i} \tag{2.32}$$

- Mean Intersection over Union

$$\frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \tag{2.33}$$

- Frequency Weighted Intersection over Union

$$(\sum_k t_k)^{-1} \sum_i \frac{t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \tag{2.34}$$

Furthermore, scientists ordinarily assess the ongoing presentation of independent route frameworks by estimating the execution time and memory utilization. Albeit these markers are normally firmly identified with equipment settings, they are likewise fundamental for model improvement.

2.8.14.2 Model Evaluation

There are two primary strategies for model assessment in AI: (I) holdout approval; (ii) and cross-approval. The two techniques utilize an isolated test set of inconspicuous information in the model exhibition assessment process. While in model preparing, the goal is to limit the preparation blunder dependent on the picked metric.

- **Hold-Out Validation** It is known as a train/test-split strategy which requires a piece of the first information to be held out from the preparation cycle. The last assessment score is just figured by trying different things with the test set on the created model. This strategy is basic, moderately quick, and guarantees that the model is tried on concealed information. The primary hindrance of this strategy is that a piece of the first information is eliminated from the preparation set of the model. Besides, there is a danger to have a high fluctuation in the forecasts.

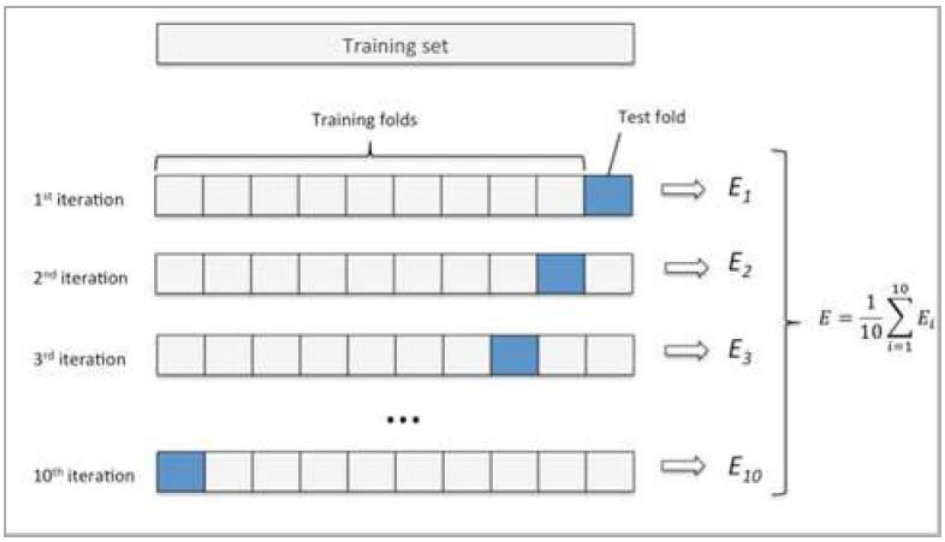


Figure 2.17: 10-crease cross-approval. The assigned preparing set is additionally split into K folds (K=10), every one of these will currently work as a hold-out test set in K emphasess. At long last, the scores got from the model on individual cycles are added and found the middle value of into the last score

- **Cross-Validation** (182, 183). For every cycle, $\frac{(k-1)}{k}$ proportion of the first information is utilized for preparing a model, while the excess parcel with a size of $\frac{1}{k}$ is utilized for

model approval. With every cycle, the unused segment is saved for the approval test, and another model is prepared from the leftover segments. The last score is registered by averaging over the quantity of cycles, k.

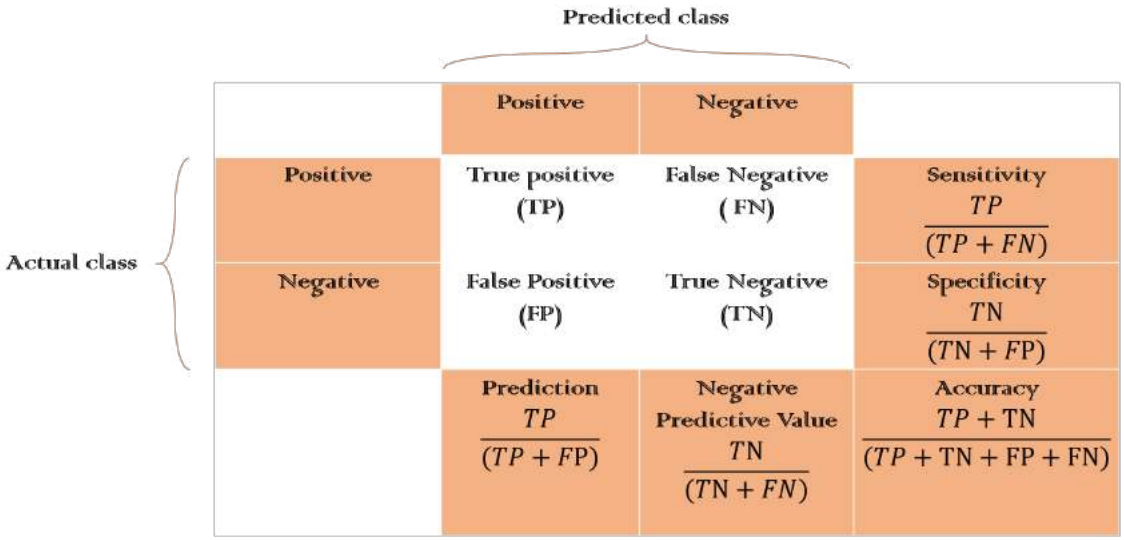


Figure 2.18: An illustrative portrayal of the (double) disarray framework and a choice of the actions that might be gotten straightforwardly from it

To assess AI techniques and contrast them and different works, scientists ordinarily utilize the most well-known arrangement models assessment measurements, which include:

- **Confusion Matrix**: A disarray network gives an extremely instinctive and complete outline of grouping models' presentation. The network has an element of $N \times N$, where N is the quantity of target class names of the thought about issue. The ground-truth (genuine name of cases) is coordinated with the forecasts came about because of the prepared model, showing data on how much the model is precise in the expectations for each class with giving the circulation of misclassified occasions in each class. The disarray lattice is the reason for figuring most assessment measurements in AI. In a twofold (two-class) model 2.18, four essential counts are gotten from the grid: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). In view of these terms, we can infer the majority of the measurements portrayed underneath.

- **Accuracy**: The most took on and refered to execution metric, characterized as the level of accurately arranged models (occasions) out of the complete number of models. Precision (ACC) is a decent metric when the class appropriation is adjusted. The issue of imbalanced class dissemination becomes clear when one class overwhelms another

class(es). For instance, in a dataset of 900:100 (class 0:class 1) paired class conveyance, ordering indiscriminately all cases as regrettable will result in 90% of accuracy.

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{2.35}$$

- **Precision and Recall**: Precision is characterized as the proportion of effectively anticipated names out of the absolute anticipated positive names. Review is characterized as the proportion of effectively anticipated encouraging points to the absolute number of positive names (ground-truth) in the information.

$$Accuracy = \frac{TP}{TP + FP} \tag{2.36}$$

$$Review = \frac{TP}{TP + FN} \tag{2.37}$$

- **F-measure**: It is known as the decent F-measure. It is a solitary scalar worth metric rundown that consolidates both Precision and Recall measurements together. It is utilized in the presentation assessment of double grouping issues. F1 is characterized as the consonant mean of accuracy and review, which similarly gauges accuracy and review. As it depends on review and accuracy, the F-score just thinks about the positive expectations. F1-score is for the most part viewed as in the assessment when class irregularity is an issue (184). A high F-measure score is a decent mark of a decent performing classifier w.r.t. minority classes.

$$F\_Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.38}$$

2.9 Related work in Deep learning applied to CBIR techniques

In this part, we expect to introduce a complete audit of the new improvement in the space of CBIR and picture portrayal. We investigated the fundamental parts of different picture recovery and picture portrayal models from low-level component extraction to ongoing semantic profound learning draws near.

2.9.1 Low-Level Feature Fusion

Ashraf et al. (185) introduced a CBIR model that depends on shading and discrete wavelet change (DWT). For the recovery of comparative pictures, the low-level element tone, surface, and shape are utilized. These highlights assume a critical part in the recovery interaction. Various sorts of elements and element extraction method are examined and situations in

which highlight extraction procedure is great are clarified (185). To set up the eigenvector data from the picture (185), shading edge discovery and discrete wavelet approaches are utilized. The shading space RGB and YCbCr are utilized to separate the shading highlights. The specialists in (185) changed RGB pictures to YCbCr shading space to extricate the significant data. To recover the question picture, the shading and edge-based highlights are extricated to register the component vector. Assuming there is a little distance between the inquiry picture and vault picture, the related picture from the information base is chosen to coordinate with the picture that is passed in question. To lessen the computational advances and improve the pursuit, the shading highlights are likewise fused with histogram and the Haar Wavelet change was applied. And afterward for picture recovery, the fake neural organization (ANN) is applied; then, at that point, its presentation is estimated against the current CBIR framework. the outcome shows that this strategy has a preferable exhibition over the others (185).

Ashraf et al. (186) introduced another CBIR method that utilizes the mix of shading and surface highlights to extricate the nearby vector which is utilized as an included vector. Shading minutes are utilized to separate the shading highlight, and for the surface component, the discrete wavelet change and Gabor wavelet strategies are utilized. To improve the element vector tone and edge, registry descriptor is likewise utilized in the component vector. then, at that point, this strategy is contrasted and any remaining existing CBIR strategies and great execution is accomplished (186) as far as accuracy and review esteems.

Mistry et al. (187) led a concentrate on CBIR by utilizing mixture highlights and different distance measurements. Mixture highlights consolidate three unique elements descriptors which comprise of spatial elements, recurrence, binarized measurable picture highlights (BSIF), and shading and edge directivity descriptors (CEDD). Highlights are separated by utilizing BSIF, CEDD, HSV shading histogram, and shading second. Highlights that are removed by utilizing HSV histogram contain shading quantization and shading space change and histogram calculation. Various examinations are performed on that methodology, and the outcomes show that this methodology fundamentally performs better compared to the current techniques (187) .

Ahmed et al. (188) led a concentrate on CBIR by utilizing picture highlight data combination. In this procedure, the combination between the removed spatial shading highlights with shape highlights extricated and object acknowledgment happens. Colors with shape together can separate the article all the more precisely. Spatial shading highlight in the element vector builds the recovery of the picture. In the proposed strategy, RGB tone is utilized to extricate the shading highlight while the graylevel pictures are utilized to separate the article edges and corner in the development of shape. the recognition of corner and edges from

the shape makes all the more remarkable descriptor. Shape recognition adjusts the better comprehension of article or picture. Shape picture recognition based on edges and corner arrangement joining with the shading produces more precise outcome for recovery or location of picture. For choosing the high difference part, the aspect decrease happens on the component vector. then, at that point, the minimal information highlights are the contribution of Bag of Word (BoW) for fast ordering or recovery of picture. the aftereffects of the trial performed dependent on this method show that it beats the current CBIR strategy (188).

Liu et al. (189) proposed a strategy for ordering and looking through a picture by intertwining the neighborhood base example (LBP) and shading data highlight (CIF). For determining the picture descriptor, the LBP removes the textural highlight. Be that as it may, the LBP has bad execution for the shading highlight descriptor.

Both the shading highlight and textural include are utilized for the effective recovery of the shading picture from a huge arrangement of data sets. In this proposed technique, another shading highlight CIF with the LBP-based component is utilized for picture recovery just as for arrangement. CIF and LBP both together address the shading and textural data of a picture. A few trials are performed utilizing a huge arrangement of data sets, and the outcomes show that this strategy has great execution for recovery and grouping of the pictures (189).

Zhou et al. (190) led a review on community oriented record implanting. This work investigates the capability of binding together ordering of SIFT include and the profound convolutional neuron organization (d-CNN) for the recovery of picture. To check the common picture level area structure and to verifiably incorporate the CNN and SIFT highlights, file the community oriented record inserting calculations proposed which persistently update the file document of CNN and SIFT highlights. After constant emphasis of the installing file, the CNN implanted record is utilized for the internet based inquiry, which shows the effective recovery precision with 10

Li et al. (191) contemplated on the shading surface component picture which depends on the Gaussian copula model of Gabor wavelets. He proposed a productive strategy for the recovery of the picture in the shading and surface setting by utilizing the Gaussian copula model which depends on Gabor wavelets. Gabor channel is considered as a straight channel which is utilized for signal examination. Direction and the recurrence portrayal of Gabor channel are taken after with the human visual framework and it is especially utilized for surface picture recovery and the copula model is utilized to catch the reliance structure in the variable where conditions exist. Gabor wavelets are utilized to deteriorate the shading picture; after decay, three kinds of conditions exist in disintegrated subbands of Gabor wavelet. these three conditions are directional reliance, shading reliance, and scale reliance. After the decay,

presence conditions are dissected and caught by utilizing the Gaussian copula strategy. there are three kinds of plans created for Gaussian copula, and as needs be, four Kullback–Leibler distances (KLD) are presented for shading recovery picture.

A few examinations are performed utilizing the datasets A LOT and STex, and the outcomes show that it performs better compared to the few cutting edge recovery strategies (191). Bu et al. (192) concentrated on CBIR by utilizing shading and surface highlights by consolidating the shading and surface elements extricated from the picture utilizing Multi-Resolution Multi-Directional (MRMD) channels. MRMD channels are utilized as basic and it tends to be free to low-and high-recurrence highlights, and it produces effective multiresolution multidirectional examinations. HSV shading space is utilized as its qualities are exceptionally near the human visual framework. Neighborhood and worldwide elements are extricated from the area of low-and high-recurrence in each shading space. A few investigations are performed by contrasting the accuracy VS review of the recovery and the element aspect vector. the outcomes show that this strategy has critical improvement over the current procedures (192). A point by point rundown of the previously mentioned low-level element combination for CBIR is addressed in Table 2.2 and 2.3.

Nazir et al. (193) led a concentrate on CBIR by combining the shading and surface elements. Since recovering the picture from an enormous arrangement of information bases is a difficult undertaking, analysts proposed numerous methods to defeat this test. Nazir et al. (193) utilized both the shading and surface highlights to recover the picture. the past research shows that by recovering the picture utilizing a solitary component doesn't give great outcomes and utilizing various highlights for picture recovery is by all accounts a superior choice. the shading highlight is removed utilizing the shading histogram while the surface component is extricated utilizing discrete wavelet change (DWT) and by edge histogram descriptor. In the extraction of shading highlights, the shading space of the picture portrays the shading cluster. HSV shading space is utilized for shading highlight, as announced the tone and immersion is exceptionally near the human visual framework. the DWTis utilized for surface component extraction since it is extremely proficient for nonstationary signal. It differs for both the recurrence and spatial reach. Here, the creator applied "Daubechies dbl" wave as it gives extremely effective outcome than the others. Edge histogram descriptor is utilized to portray just the conveyance of neighborhood edges in the picture. EDH is utilized to observe the most applicable picture from the data set and it played out some computational advances, and finally, EDH is determined for the picture. Various tests are utilized to decide this strategy; accordingly, it performs better compared to the current CBIR framework (193).

2.9.2 Local Feature-Based Approaches

Kang et al. (194) directed a review on picture likeness appraisal strategy dependent on meager element portrayal. To naturally decipher the comparative things in various pictures is the principle purpose for likeness appraisal. Data loyalty issue is taken as the picture comparability evaluation issue. For social affair data accessible in the reference picture and assessing the measure of data that can be gathered from the test picture, an element based methodology is proposed (194).

Zhao et al. (195) proposed agreeable meager portrayal in two inverse ways for semi-directed picture comment. Inadequate portrayal is successful for some PC vision issues and its bit form has incredible order ability. They zeroed in on agreeable SR application in the semi-regulated picture explanation which might build the quantity of named pictures in the preparation picture classifiers for sometime later. A bunch of named and unlabeled pictures is given, and the typical SR system which is otherwise called forward SR is utilized to address each unlabeled picture with numerous other marked pictures, and from that point onward, the unlabeled picture is explained by the name picture comments. In reverse SR approach, the explanation cycle is finished and names are appointed to the pictures that are without semantic portrayal. the primary spotlight is on the commitment of in reverse SR to picture explanation.

thiagarajan et al. (196) led a review on managed nearby scanty coding of subimage highlights for picture recovery. In the wake of being broadly utilized in picture demonstrating, inadequate portrayal is presently being utilized in utilizations of PC vision. the highlights that separate one picture from the other should be extricated for recovering and grouping pictures. To perform directed nearby meager coding of bigger covering districts, an element extraction approach is proposed which utilizes different worldwide/neighborhood highlights. A technique is proposed for planning word reference and regulated neighborhood inadequate coding of sub-picture heterogeneous highlights. Exploratory outcomes show that proposed highlights beat the spatial pyramid highlights acquired utilizing neighborhood descriptors.

Hong and Zhu (197) proposed a clever positioning strategy with QBME for recovering pictures quicker which depends on an original learning structure. the current QBME approach utilizes all models separately and afterward consolidates their outcomes wherein on every addition of inquiry model their computational time additionally increments. To begin with, the semantic relationship, which is mastered utilizing meager portrayal, of picture information in the preparation cycle is investigated. A semantic connection hypergraph is built to display the connection between pictures in the dataset. A pre-learned semantic relationship is utilized subsequent to building SCHG to assess the connecting esteem among pictures. Second, a various examining procedure is proposed to rank the pictures with numerous question

models. The current QBME strategy acknowledges each info model in turn, yet in the proposed technique, all information models are handled simultaneously. along these lines, the proposed conspire shows viability as far as speed and recovery execution.

Wang et al. (198) completed a review on recovery based face comment by frail name regularized neighborhood organize coding. To identify a human face from a picture and clarify it as per the picture naturally is critical to some genuine applications. A system is given to resolve the issues in mining monstrous web facial pictures accessible on the web. For a given question picture, first utilizing content-based picture recovery, top "n" pictures from web facial picture information bases are recovered and afterward their names are utilized for auto comments. This technique has two primary issues that are (1) how to coordinate with the question picture and pictures put in the chronicle and (2) how comparable names can be doled out to the pictures that are not corresponded with one another. A WLRLCC method is proposed which takes advantage of the standard of both neighborhood facilitate coding and chart based frail name association.

Srinivas et al. (199) did a review on content-based clinical picture recovery sing word reference learning. For gathering huge clinical datasets, a bunching strategy utilizing word reference learning is proposed. A K-SVD bunches comparable pictures into the groups utilizing word references. A symmetrical coordinating with pursuit (OMP) calculation is utilized to coordinate with an inquiry picture with the current word reference to distinguish the word reference with the sparsest portrayal. For recovering the pictures that are like the question pictures, the pictures remembered for the bunch related with this word reference are analyzed utilizing likeness measure. the best thing about this methodology is that it doesn't need preparing and functions admirably on various clinical data sets. A pictures information base named IRNA is utilized for assessing the presentation of the proposed technique. Results exhibit that the proposed strategy proficiently recovers picture from clinical information bases.

Mohamadzadeh and Farsi (200) directed a review on content-based picture recovery framework through scanty portrayal. A few media data handling frameworks and applications require picture recovery which observes inquiry picture in picture datasets and afterward addresses as required. Concentrates on show that the pictures are recovered in two ways, i.e., text-based and content-based picture recovery. the reason for the recovery frameworks is to recover the picture consequently as indicated by the question. However, numerous analysts are drawn in towards the speed and exactness with which the pictures are recovered consequently. the proposed conspire utilizes scanty portrayal to recover pictures.

Li et al. (201) proposed a clever sketch-based imaged recovery utilizing item quantization with meager coding to build the codebook. In this technique, the ideal picture sketch is

drawn and includes are separated utilizing the cutting edge neighborhood descriptors. then, at that point, by utilizing item quantization and inadequate coding, writers (201) encoded the elements into the advanced codebook and afterward encode the sketch highlights utilizing quantization remaining to further develop the portrayal capacity. Consequently, this technique can be effectively registered and great execution is accomplished contrasted with a few famous SBIR. Because of the item quantization, its advantage is that it tends to be immediately executed.

Picture recovery is a method to peruse, search, and recover the picture for a huge arrangement of information base. It gives accommodation to living souls (202). AI is adequately expanding the nature of recovery. AI is additionally effectively utilized for picture explanation, picture characterization, and picture acknowledgment. Various procedures are utilized to recover the picture utilizing shading and surface elements. It is hard for basic component extraction procedure to get the significant level semantics data of target data; thus, for this arrangement, various models are proposed which add to extricate the semantic data of the objective picture. Because of headway in AI, profound learning has showed up in many fields

of current life. In the profound adapting additionally, various strategies are introduced. It is to be notice that the scanty portrayal model depends on the establishment of inadequate portrayal. In any case, the excellent of the picture recovery result is gotten from an enormous number of learning occasions. In any case, with the wastage of numerous HR, it additionally possesses a lot registering assets. To tackle this issue, the creators proposed the inadequate coding-based not many learning occurrences model for picture recovery.

As indicated by Duan et al. (203), face acknowledgment acquired high consideration in PC vision. Over the most recent twenty years, many face acknowledgment techniques are presented. there are two fundamental methods for face acknowledgment: one is to extricate the discriminative component from the face so it can isolate face picture of various individual and the second is that the face coordinating is to plan successful classifiers to perceive distinctive individual. Countless face acknowledgment strategies are proposed over the most recent couple of years, which are chiefly arranged into all encompassing and neighborhood highlight portrayal.

An itemized rundown of the previously mentioned nearby element for CBIR is addressed in Table 2.2 and 2.3. Figure 2.18 gives an outline of ordinarily utilized methods of AI for CBIR system and Figure 2.20 is about the vital disciplines of machine-human cooperations. Histogram-based picture portrayal separates nearby elements and afterward encodes them. This cycle requires a precomputed codebook, otherwise called visual jargon. Assuming that there are n quantities of picture datasets, separate codebook is needed to be figured for each case and this interaction requires high computational expense (204). If there should be an

occurrence of a predetermined number of preparing tests, the processed codebook can be one-sided and it can corrupt the exhibition of the BoVW model. When the precomputed codebook from any dataset is applied for on the web/new arrangement of pictures, the segregating capacity of codebook diminishes (204).

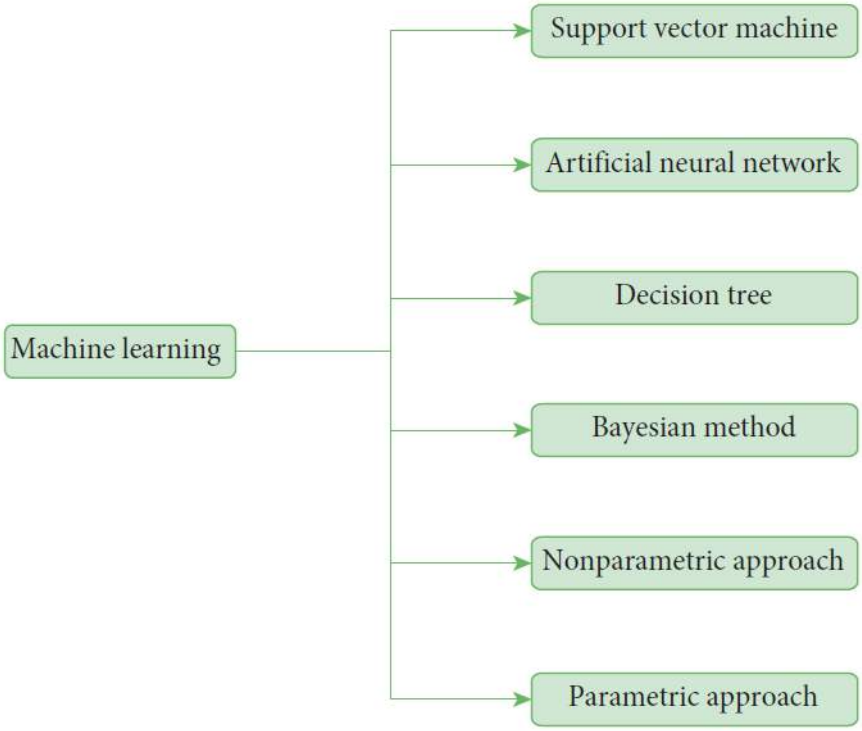


Figure 2.19: ML techniques on CBIR

To beat this constraint, the writers proposed a novel verifiable codebook move technique for visual portrayal (204). the proposed approach is unique in relation to the past research as it depends on a prelearned codebooks dependent on nonlinear exchange. For this situation, the neighborhood highlights are recreated based on nonlinear change and understood change is conceivable. This methodology gives the utilization of prelearned codebooks for new visual applications through understood learning. the proposed research is approved through a few standard picture benchmarks, and exploratory outcomes exhibit the adequacy and productivity of this understood learning (204).

The writers (205) proposed a clever fine-grained picture order model by utilizing a mix of codebook age with low-rank meager coding (LRSC). Classspecific and conventional codebooks are figured by applying improvement on aggregate recreation blunder, the sparsity imperatives, and confusion of codebook. As per (206), picture visual highlights assume a crucial part in independent picture grouping. Nonetheless, in PC vision applications, the presence of similar view in the pictures of various classes regularly brings about visual el-

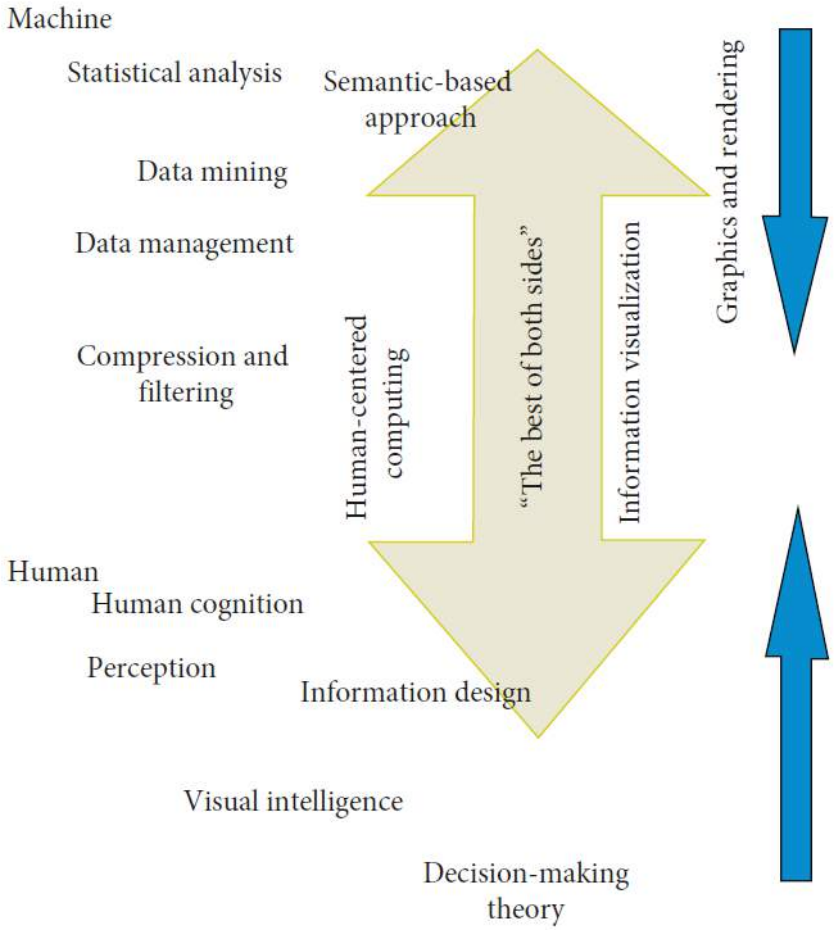


Figure 2.20: Machine-human interations

ements conflictingly. the development of unequivocal semantic space is an open PC vision research issue. To manage visual highlights conflictingly and development of express semantic space, the creators proposed organized powerless semantic space for picture grouping issue (206).

As per (207), object-driven based order for picture characterization is more dependable when contrasted with the methodologies that depend on division of the picture into subregions like SPM. To observe the area of an item inside the picture is an open issue for PC vision research local area. As indicated by (207), the presentation

of picture arrangement model debases in case the accessible semantic data inside the picture is disregarded. the creators proposed an original methodology for object classification through Semantically demonstrating the Object and Context data (SOC). A prelearned classifier is applied by figuring relationships of every applicant district with high certainty scores, and these areas are assembled as a bunch for object choice. different spaces of the pictures in which there is no article are treated as the foundation. This methodology gives

an interesting and discriminative component for object classification and portrayal (207). As per (208), managed learning is for the most part utilized for arrangement and grouping of advanced pictures. Directed learning is reliant upon named datasets, and now and again, when there are an excessive number of pictures, it is hard to deal with the naming system.

2.9.3 CBIR Research Using Deep-Learning Techniques

Looking for computerized pictures from ale stockpiling or data sets is frequently required, so content-based picture recovery (CBIR) otherwise called question based picture recovery (QBIR) is utilized for picture recovery. Many methodologies are utilized to determine this issue, for example, scale-invariant change and vector of privately accumulated descriptor. Because of most conspicuous outcomes and with an incredible presentation of the profound convolutional neural organization (CNN). By furnishing amazing picture recovery procedures with a superior result, these methodologies link the TF-IDF with CNN examination for visual substance. To demonstrate the proposed model, the creators direct analysis on four picture recovery datasets and the results of the trials show the presence of the reality of the model.

Shi et al. (209) proposed a hashing calculation that concentrates highlights from pictures and learns their double portrayals. the creators model the pairwise grid and a genuine capacity with profound learning system that learns the twofold portrayals of pictures. Tests are directed on a huge number of histopathology pictures (on 5356 skeletal muscle and 2176 cellular breakdown in the lungs pictures with 4 kinds of illnesses) to show the reliability of the proposed calculation. the proficiency of the proposed calculations is accomplished with 97.94% order exactness.

Zhu et al. (210) proposed unaided visual hashing approach known as the semantics helped visual hashing (SAVH). This framework utilizes two parts that are disconnected learning and internet learning. In disconnected adapting initially, the picture pixel is changed into numerical vector portrayal by removing the visual and surface element. then, at that point, text improving the visual chart is removed with the help of subject hypergraph, and the semantics data is separated from the text data and afterward the hash code of picture is realized which saves the relationship of picture between the semantics and pictures, and afterward at the last, the hash work code is produced inside the direct forceful model. these advantageous properties match the prerequisite of genuine application situations of CBIR (210).

In PC vision applications, the utilization of CNN has shown a striking presentation, particularly in CBIR models. The vast majority of the CNN models get the highlights in the last layer utilizing a solitary CNN with request less quantization approach and its disadvantage is they limit the use of halfway convolutional layer for distinguishing neighborhood picture

design. Thus, in this paper, another procedure is distinguished as bilinear CNN-based design. This technique utilized two equal CNN models to separate the component without the earlier information on the semantics of picture content. the element is straightforwardly removed from the actuation of the convolutional layer as opposed to diminishing very lowing dimensional component. the test on this methodology gives a vital end: This model diminishes the picture portrayal to the reduced length as it utilized distinctive quantized levels to remove the component, so it is exceptional to support the recovery execution and the hunt time and capacity cost. Also, the bilinear CRB-CNN is exceptionally powerful in learning an extremely mind boggling picture having diverse semantics. Ten milliseconds is expected to extricate the component from the picture and search from the information base and tiny circle size is expected to address and store the picture. What's more, toward the end, start to finish tanning is applied with next to no other metadata, comments, labels which adjusted the ability of CRB-CNN to extricate the component from just the visual data in CBIR task. This strategy likewise applies for the huge scope information base picture to recover the picture and showed a high recovery execution (211).

For productive picture search, hashing capacity acquires effective consideration in CBIR (212). Hashing capacity gives a comparative twofold code to the comparable substance of the picture which maps the high-dimensional visual information into low-dimensional paired space. This methodology is essentially relying on the CNN. It is to be accepted that the semantic marks are addressed by the few inactive layer credits (twofold code) and characterization additionally relies on these properties. In light of this methodology, the administered profound hashing strategy builds a hash work from an inactive layer in

the profound neurons organization and the double code is gained from the true capacities that clarified about the characterization blunder and other advantageous properties in the parallel code. the principle component of the SSDH is that it brings together recovery and arrangement in a solitary model. SSDH is versatile to enormous scope search, and by slight change in the current profound organization for characterization, SSDH is basic and effectively feasible (212). A definite synopsis of the previously mentioned profound learning-based highlights for CBIR is addressed in Table 2.2 and 2.3.

Compelling picture examination and order of the visual data utilizing discriminative data is considered as an open exploration issue (213). Many exploration models are proposed utilizing various methodologies either by consolidating sees by diagram based methodology or by utilizing move learning. It is troublesome from the current strategies to process the discriminative data at the picture borders and to find closeness consistency limitation. the creators (213) proposed a multiview name sharing strategy (MVLS) for this open examination issue and attempted to keep up with and hold the likeness. For visual order and portrayal,

streamlining over the change and arrangement boundaries is consolidated for change lattice learning and classifier preparing. Results on MVLS with various six perspectives (no intra-view and no between view in addition to no intra view) and nine perspectives (mix of intra-view and between view) are directed. Trial results are contrasted and a few cutting edge exploration and results shows the adequacy of the proposed MVLS approach (213).

For the comprehension of pictures and article classification, strategies like CNN and neighborhood include have shown great execution in numerous application spaces. the utilization of CNN models is as yet trying for exact classification of item and for the situation with restricted preparing data and names. To deal with the semantic hole, the smooth requirements can be utilized, however the exhibition of the CNN model corrupts because of the more modest size of the preparation set. the creators (214) proposed a multiview calculation with few marks and view consistency (MVFL-VC). Both named and unlabeled pictures are utilized together for the picture view consistency with multiview data. the discriminative force of the learned boundary is likewise improved by unlabeled preparing pictures. To assess the proposed calculation, tests are directed on various datasets. the proposed MVFL-VC calculation can be utilized with other picture arrangement and portrayal methods. the calculation is tried on unlabeled and concealed datasets. the aftereffects of tests and investigation uncover the adequacy of the proposed technique zhang2018multiview. the extraction of area space information can be useful to diminish the semantic hole (214). the creators proposed multiview semantics portrayal (MVSR), which is a semantics portrayal for visual acknowledgment.

The proposed calculation separates the pictures based on semantic and visual likenesses (214). Two visual likenesses for preparing tests give a steady and homogenous discernment that can deal with various parcel methods and various perspectives. the proposed research dependent on MVSR is more discriminative than other semantics approaches as the semantic data is figured for later use from each view and from isolated assortment of pictures and various perspectives. Diverse openly accessible picture benchmarks are utilized to assess this exploration, and the test results show the viability of MVSR. the outcome exhibited that MVSR further developed order execution as far as accuracy for picture sets with more visual varieties.

2.10 Conclusion

In this chapter, we had introduced two parts. We talked in the first part about image representation, descriptors, and segmentation techniques, annotation techniques, search techniques, and techniques recommendation. We had introduced the second part the machine learning, classification, and optimization techniques. Based on the presented state-of-the-art annotation techniques, we may say that choosing or knowing how to combine two metrics in order

Author	Dataset	Application	Accuracy	Precision
Nazir et al. (193)	Corel 1-k		-	0.735
Ashraf et al. (186)	Corel 1000	CBIR	-	0.875
Mistry et al. (187)	Wang	CBIR	-	0.875
Ahmed et al. (188)	Corel 1000	CBIR	-	0.90
Kang et al. (193)	COIL-20	Image similarity assessment	0.985	-
Zhao et al. (195)	ImageCLEF-VCDT	Semi-supervised image annotation	-	-
Thiagarajan et al. (196)	Cambridge image dataset	Image retrieval	0.97	-
Hong and Zhu (197)	Yale face dataset	Transudative learning image retrieval	0.65	-
Wang et al. (198)	WDB and ADB	Retrieval-based face annotation	-	-
Srinivas et al. (199)	Image CLEF	Medical CBIR	0.5	-
Mohamadzadch and Farsi mohamadzadeh2016content	Flower, Corel	CBIR	-	-
Li et al. (201)	Eitz	SBIR	0.98	-
Duan et al. (203)	LFW, YTF, FERET	Face recognition	0.846	-

Table 2.2: Summury 1 of ML in CBIR

Author	Dataset	Application	Accuracy	precision
Krizhevsky et al. (215)	ILSVRC-2010 ILSVRC-2012	Image classification	37.50% top-1 and 17% to-5 error rate on ILSVRC-2010 and 15.3% top-5 error rate on ILSVRC-2012	
Sun et al. (216)	LFW (labeled face in the wild)	Face verification	97.45%	
Karpathy and Fei-Fei (217)	Flickr8K, Flickr 30 K and MSCOCO	Generation of descriptions of image regions	..	
Li et al. (218)	MIRFlickr and NUSWID	Social image understanding	CBIR 0.512 on MIRFlickr and 0.632 NUSWID with k=1000	
Kondylidis et al. (219)	INRIA Holidays, Oxford 5k, Paris 6k, UK Bench	CBIR	——	

Table 2.3: Summury 2 of ML in CBIR

to obtain a hybrid content-based image retrieval system is the key to obtaining an efficient and robust result. Due to that, we motivated by the benefits offered from the use of deep learning techniques and use them to develop a hybrid extractor and semantic interpreter of <object-relationship-subject> based images model.

3

Ontology Learning Models Overview

3.1 Introduction

Recently, several works have focused on the use of scientific advances in the field of AI to overcome the problems of the semantic gap, inter/intra classes divergence, and long-tail appearing. These systems are based on domain knowledge represented by ontologies. These provide a common, structured, and shared understanding of a domain or task, which can be used for communication between humans and machines.

Data on the Web is a great source of data represented in forms that are useful to humans but difficult for automatic processing. The content of web resources can be described using formal metadata. The latter must be based on ontologies in order to be able to exploit them by machines and provide them with semantics. This chapter is divided into two important sections, the first part will give an overview of the ontology learning models. In the second part, we will detail the used learning strategy which is the ontology learning "layer cake".

3.2 Interdependence between knowledge and language

An ontology, as a formal system of signs, fits perfectly into this presentation where it "plays the role" of natural language in information systems. Thus the semiotic triangle can be easily transformed into an "ontological" triangle: the vertices (signified, signifier, referent) is interpreted as (concept, term, instance), respectively. The relationships between the vertices do not change. In computer science, the vertices of the semiotic triangle can be associated with notions of data, information, and knowledge. Here their connection is presented differently, in the form of the "pyramid of wisdom" (see. Figure 3.2), known as DIKW (Data-Information-Knowledge-Wisdom).

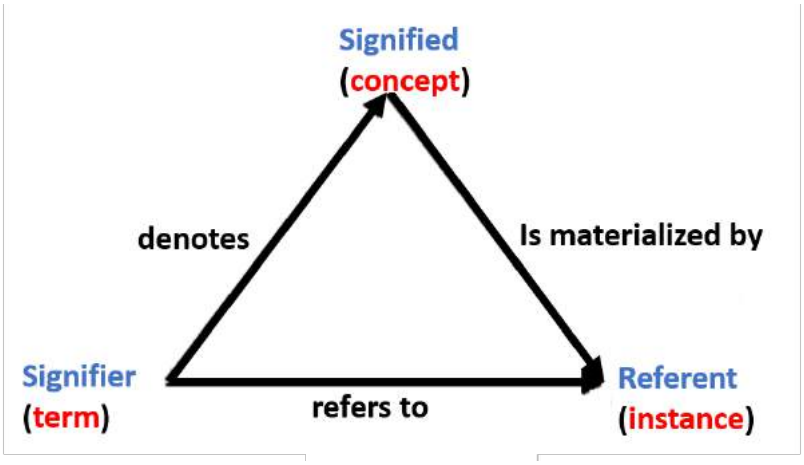


Figure 3.1: Generic semiotic triangle

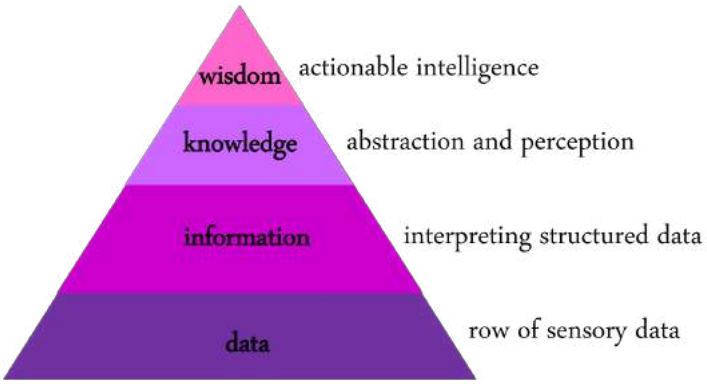


Figure 3.2: The DIKW pyramid of wisdom (220)

3.2.1 Preliminaries

In the following, we present all the definitions synthesized from the above references, by illustrating the point of view of the exact sciences, where these notions are concrete objects with measurable parameters, and that of philosophy, where the definitions rather have a character of general interpretations.

3.2.1.1 Data

- **In computer science and exact sciences**: All types of signals that can be perceived by humans or by an information system (220).

- **In philosophy and social sciences**: The concept is fuzzy; it is often interpreted as a synonym for information or facts (220).

3.2.1.2 Information

- **In computer science and exact sciences**: The quantitative magnitude characterizes the reduction of ambiguity on the choice of the state of the object among several possible variants (220).
- **In philosophy and social sciences**: The fundamental characteristic of being; the axiomatic notion such as matter, energy, space, time; it cannot be defined through the other categories. Often information is confused with facts or data, and sometimes with knowledge (220).

3.2.1.3 Knowledge

- **In computer science and exact sciences**: The result of the accumulation of skills and know-how through multiple artifacts according to authors in (221); it can be governed reliably by information systems (220).
- **In philosophy and social sciences**: The cognitive framework that enables humans to use information. Valid information, consistent with other accepted truths (220).

3.2.1.4 Concept

- **In computer science and exact sciences**: Corresponds to the class of objects whose properties are explicitly restricted by the imposition of constraints. A concept corresponds to an intentional definition of the class (set) of its referents (222).
- **In philosophy and social sciences**: The element of thought relating to a plurality of distinct things responding to the same law (222).

Table 3.1 is used to visualize the overlap of properties shared by the concepts of data, information, and knowledge, based on the definitions presented below. Thus, the data can be recorded or "recognized" by an information system (it is said that it is recognizable). Relevant information can be distinguished from noise: it is interpretable. Finally, knowledge, in the context of data, assumes the ability to use information, to deduce new facts that are not explicitly present (in this sense, we say that knowledge is predictive) (220).

3.3 The notion of ontology

3.3.1 Onset of ontology

Knowledge engineering (KI) is aimed at automatic solving problems, while Knowledge-Based Systems (KBS) should allow knowledge storage and consultation and modification, as well

	Recognizable	Interpretable	Predictive
Data	×	×	
Information	×	×	
Knowledge		×	×

Table 3.1: Properties that allow to distinguish data, information and knowledge (220)

as automatic reasoning on it, (223).

The sharing of knowledge between computer systems will allow interaction and cooperation between them and also between human users. This is manifested, for example, in decision support systems, computer-assisted education systems, searching for information on the web, etc. To allow efficient automatic processing, the representations to be used by the machines (models) must be loaded with meaning, and this by linking the information that is collected and represented to other types of information essentially devoted to the underlying semantics. This gave birth to ontological engineering.

The term "Ontology" is a philosophical term that means "Part of metaphysics which applies to being as being, regardless of its particular determinations" (224). With the emergence of knowledge engineering (KI), ontology is introduced in Artificial Intelligence (AI) as a response to the problems of representation and manipulation of knowledge within computer systems.

3.3.2 Definition

It is hard to give the thought of metaphysics a solitary conclusive definition since it has been utilized in various settings. Creators in (225) were quick to propose a meaning of cosmology, in particular: "a metaphysics characterizes the terms and the essential relations of the jargon of an area just as the guidelines which demonstrate how to consolidate terms and connections with the goal that the jargon can be broadened".

Creators in (25) characterized philosophy as "an express particular of a conceptualization". This definition has been adjusted somewhat by creators in (226) as "formal particular of a common conceptualization". These two definitions have been consolidated by (227) as a "formal and express detail of a common conceptualization":

- **"Conceptualization"** alludes to a theoretical model of specific peculiarities on the planet, a model which recognizes the important ideas of this peculiarity.
- **"Explicit"** implies that the sort of ideas utilized and the limitations on their utilization are expressly characterized.

- **"Formal"** alludes to the way that metaphysics should be reasonable by machines.
- **"Shared"** mirrors the thought of consensual information depicted by cosmology, in other words, that it isn't confined to the perspective of specific people just, yet mirrors a perspective. more broad view, shared and acknowledged by a gathering.

As per creators in (228), a philosophy is "a bunch of terms organized in a various leveled style, intended to portray a space and which can fill in as a structure for an information base". Creators in (229) referenced that the motivation behind a metaphysics is the investigation of classes of ideas that exist or may exist in specific spaces. The consequence of this review, called a cosmology, is a list of sorts of things that exist in a space of interest with the possibility of utilizing a proper language to talk about an area.

Note that the definitions, notwithstanding their variety, offer corresponding perspectives. Thus, an ontology offers the means to present the concepts of a domain by organizing them hierarchically and by defining their semantic properties in a formal knowledge representation language, the following figure is a Hierarchy of concepts of a "pizza" ontology captured by Protégé [1].

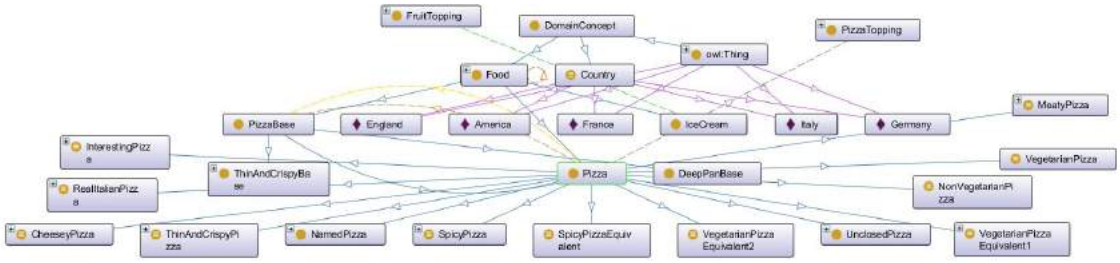


Figure 3.3: Hierarchy of concepts of a "pizza" ontology

3.3.3 Ontology Types

We can define two classes of ontologies: (i) a category based on the structure of the conceptualization and (ii) a second category based on the subject of the conceptualization (230). In the first category, there are three subcategories:

- Terminological ontologies (lexicons, glossaries…).
- Information ontologies (DB diagram).
- The ontologies of knowledge models.

[1] http://protege.stanford.edu/ontologies/pizza/pizza.owl

In the second category, there are four subcategories:

- **Application ontologies:** they contain all the data expected to show information for a specific application.
- **Domain ontologies:** they give a bunch of ideas and connections depicting the information on a particular space.
- **Generic ontologies (likewise called significant level ontologies):** they are like area ontologies, yet the characterized ideas, are more conventional and depict information like state, activity, space, and parts. Normally, the ideas of a space metaphysics are specializations of the ideas of an undeniable level cosmology.
- **Representation ontologies (meta-ontologies):** they give formalization natives to the portrayal of information. They are by and large used to compose space ontologies and significant level ontologies. Models: Frame Ontology and RDF Schema Ontology.

3.3.4 Ontology Component

An ontology is made up of several components namely: the concepts often represented by terms, the relations between these concepts, the axioms, and the instances.

3.3.4.1 Concepts

As per creators in (231), an idea addresses a material item, quality, or thought. It is comprised of three sections: at least one terms, an idea, and a bunch of items:

- **The term** additionally called name, permits to assign the idea.
- **The attribute** additionally called the intension of the idea, compares to the semantics of the idea, communicated as far as properties.
- **The set of objects** additionally called an expansion of the idea, assembles the items controlled through the idea; These articles are called examples of the idea.

Authors in (232) identified the properties that a concept of ontology can have:

- **Genericity:** a concept is generic if it does not have an extension.
- **Identity:** a concept carries an identity property if this property allows one to conclude as to the identity of two instances of this concept.

This property can relate to attributes of the concept or to other concepts.

- **Rigidity:** a concept is said to be rigid if any instance of this concept remains an instance in all possible ways.

- **Anti-rigidity:** a concept is anti-rigid if any instance of that concept is essentially defined by its belonging to the extension of another concept.

- **Unity:** a concept is a unit concept, if for each of its instances, the different parts of the instance are linked by a relationship that does not link other instances of the concept.

Likewise, two concepts can have the following properties:

- **Equivalence:** two concepts are equivalent if they have the same extension.

- **The disjunction:** two concepts are disjoint if their extensions are disjoint.

- **The dependency:** An idea C_1 is subject to an idea C_2 if for any example of C_1 there exists an occasion of C_2 which is neither part nor constituent of the case of C_1. Model: parent is an idea subject to kid (as well as the other way around).

3.3.4.2 Relations

Relationships represent a type of interaction between concepts in a certain field. They generally break down into two categories: taxonomic relationships and associative or non-taxonomic relationships. **Taxonomic relations** organize all the concepts of ontology in a tree structure. According to authors in (233), a C_1 concept subsumes a C_2 concept if any semantic property of C_1 is also a semantic property of C_2, in other words, C_1 is more specific than C_2. **Associative relationships** are interaction relationships between two concepts that are non-taxonomic. Examples of non-taxonomic relationships are "associated-a", "is-located-in", "adjacent to", "disconnected-to".

The relations of a philosophy assign various collaborations and connections between's the ideas of the metaphysics. These relations unite the accompanying affiliations: subclass of (particular or speculation), part of (conglomeration or piece), related a, an example of, is a… and so forth **Generalization/specification** (speculation is a backwards connection of specialization) is a double connection between an overall idea and a more explicit idea. **Composition/aggregation** (compound-part relationship) is a relationship that permits you to construct complex ideas from different ideas called parts.

3.3.4.3 Axiomes

The aim of axioms is to define in a logical language, the description of the concepts and the relations that represent their semantics. They represent the intentions of the concepts and relations of a certain field (234). The joining of maxims in a philosophy can have a few targets:

- Define the importance of parts.
- Define limitations on the worth of qualities.
- Define the contentions of a connection.
- Check the legitimacy of the predetermined data or gather new ones.

3.3.4.4 Instances

Instances comprise the extensional meaning of philosophy and pass on information (static, verifiable) about the field of the issue. The individual is a case of an idea, all in all, it is the component depicted by the idea, for instance, the people $Djamel$ and $Samir$ are cases of the idea $Person$.

3.3.4.5 Functions

They establish uncommon instances of connection in which a component of the connection, the n^{th}, is characterized by the $n-1$ going before components.

3.3.5 Classifications of ontologies

We can recognize four typologies as per a few models (235):

- Classification as per the degree of fulfillment.
- Classification as per the object of conceptualization.
- Classification as per the degree of formalism.
- Classification as per the degree of detail.

3.3.5.1 Classification according to the object of conceptualization

Ontologies can be subdivided into several levels which are, among others:

- **Thesaurus type ontologies:** Are also called taxonomy. They are used to define a reference vocabulary.
- **Domain ontologies:** These ontologies express domain-specific conceptualizations. They are reusable for several applications in this field. Domain ontology characterizes knowledge of the domain where the task is performed.
- **Applicative ontologies:** These ontologies contain knowledge of the domain necessary for a given application, they are specific and cannot be reused.
- **Generic ontologies or undeniable level ontologies (top-ontologies):** These ontologies express substantial conceptualizations in various field. Its subject is the investigation of the classes of things that exist on the planet. Like ideas of high deliberation like elements, occasions, states, activities, time, space, connections, and so on
- **Representation ontologies or meta-ontologies:** These ontologies conceptualize the natives of information portrayal dialects.
- **Geographical ontologies:** The ontologies of room all the more explicitly committed to the depiction of ideas that describe space like point, line, and so on These ontologies are normally evolved by huge associations of normalization.
- **Spatialized (or Spatio-temporal)**: ontologies will be ontologies whose ideas are confined in space. A worldly part is regularly important as a supplement for the demonstrating of geographic data.

3.3.5.2 Classification according to the level of completeness

We can define three levels of completeness:

- **Semantic level**: All concepts, characterized by a term / label, must respect the four differential principles:
  - Community with the ancestor.
  - Difference, specification, from the ancestor.
  - Community with sibling concepts, located at the same level.
  - Difference from sibling concepts.

These principles correspond to the semantic commitment and ensure that each concept will have an associated unambiguous and non-contextual meaning. Two concepts are identical if the interpretation of the term/wording through the four differential principles results in an equivalent meaning.

- **Level-Referential:** Referential or formal concepts are characterized by a term/label whose semantics are defined by an extension of objects. Ontological engagement specifies the domain objects that can be associated with the concept, in accordance with its formal meaning. Two formal concepts will be identical if they have the same extension.
- **Operational-Level:** The concepts of the operational or computational level are characterized by the operations that can be applied to them to generate interfaces or computational engagement.

3.3.5.3 Classification by the level of detail

We can differentiate ontologies according to the level of description used:

- **Fine granularity:** This level corresponds to very detailed ontologies, thus possessing a richer vocabulary capable of ensuring a detailed description of the pertinent ideas of a space or of an assignment.
- **Large granularity:** This level corresponds to less detailed vocabularies. For example, specific use cases or users already agree in advance about an underlying conceptualization. High-level ontologies have a large granularity since the concepts they translate are normally refined later in other domains or application ontologies.

3.3.5.4 Classification according to the formalism used

We can distinguish ontologies according to the formalism used to express them.

- **Informal:** the philosophy is communicated in normal language. This can make the ontology more understandable to the user, but it can make it more difficult to verify that there are no redundancies or contradictions.
- **Semi-informal:** the philosophy is communicated in a limited and organized type of regular language; this builds the clearness of the metaphysics while lessening uncertainty.
- **Semi-formal:** the metaphysics is communicated in a fake language characterized officially.

- **Formal:** the metaphysics is communicated in a counterfeit language with formal semantics, allowing to prove properties of this ontology. The advantage of a formal ontology is the possibility of performing verification on the ontology: completeness, non-redundancy, consistency, consistency, etc.

3.3.6 Fields of application of ontologies

In this segment, we will momentarily examine the fields where ontologies are required, and the fields that apply the methodological principles underlying their construction. The diversity of possible applications shows the topicality of the problem of the automated construction of ontologies.

3.3.6.1 Semantic Web

Today, the Semantic Web is the largest area of application of semantic technologies. Its basic idea, formulated by authors in (236), is to support the extension and long-term growth of the current Web within the framework of the recommendations of W3C [1]. Today, in order to guarantee interoperability on the web, the aim is to supplement textual resources with information allowing their unambiguous interpretation by people and computers. Since 2001, many tools, techniques, and languages for describing ontologies have appeared.

3.3.6.2 Information Retrieval (IR)

This area covers activities such as:

- Finds documents that contain relevant information, corresponding to user requests. Most search engines index texts using the vector model where each text is presented as "Bag of words". The main disadvantages of this approach are:

  1. Index redundancy, the same item being denominated by different words.
  2. The words in a document are considered independent, which obviously does not correspond to reality.
  3. The words are polysemous, which induces ambiguities and leads to irrelevant results for the user.

  These drawbacks can be overcome by using conceptual indexing using domain ontologies; the concepts are then associated with the corresponding terms and linked by predefined relationships.

[1] World Wide Web Consortium

- Document classification, i.e. assigning each document to one of the predefined categories.
- Semantic clustering, i.e. clustering of documents whose subjects are similar. In this case, the main idea is the definition of the context around which the documents must be gathered. Topic extraction is one of the current tasks of knowledge engineering that have common methods with ontology learning (237).
- Production of automatic summaries.

3.3.6.3 Question-answer systems

In this case, it is about the development of interactive systems of a Question-Answer type so that the user obtains a concrete result and not just a list of references of the documents that correspond to his request question.

3.3.6.4 Integration of heterogeneous databases

The integration of heterogeneous databases is a complex issue that has become crucial in providing users with a unified interface allowing access (through queries) to heterogeneous resources. In this case, ontologies are used to specify the content of heterogeneous resources.

3.3.6.5 Software engineering

The principle of conceptualizing and distinguishing objects according to their properties is universal; it is also used in software engineering (SE). For at least 20 years, the trend has been towards the unification and specification of processes throughout the life cycle of an Information System. These are the Model-Driven Develop (MDD) strategy and the Model Driven Architecture (MDA) which are model-based software construction architecture. The MDD strategy saves time and resources, reduces workloads, and guarantees flexibility in the implementation, maintenance, testing, and simulation processes and IS interoperability through module designs and Automatically readable data models (238, 239).

3.3.7 Lexical resources

Three types of lexical resources are increasingly used in NLP methods providing solutions to the word ambiguity problem: dictionaries, annotated corpora and lexical databases. In this section, we will briefly describe the particularities of each type of tool and present several projects that are both typical and significant:(Note that, we intend in our work to use methods that are originally dedicated to NLP and we will normalize them in order to fit our objectives, i.e?, using them in CBIR model. Also, it is important to mentions that the

lexical dataset can be used in the CBIR model as a resource of information that describes the retrieved objects.)

3.3.7.1 FrameNet

FrameNet is a human-readable and machine-readable English language resource. Its creation began in 1997 as part of a large semantic lexicon project, created under the leadership of Ch. Fillmore (240) and implementing his conception of semantic frameworks. FrameNet aims at the description of the semantic and syntactic compatibility of words according to their valences which, in turn, depend on the contextual meaning of the word. The content of FrameNet continues to grow. We present several recent figures to show its magnitude: there are more than 1,100 hierarchical lexical frames. The index contains over 13,400 lexical units (LU) illustrated by annotated textual examples. In total, there are over 28,000 annotated complete text sets and over 227,000 textual sets.

FrameNet can be seen as an example of a linguistic ontology of standardized situations whose concepts are realized in the form of frames linked by hierarchical relationships. FrameNet utilizes eight sorts of connections between outlines 8 which are gathered into three gatherings: speculation connections, event structure relationships, and "Systematic" relationships (241).

The most frequent relationships between executives are:

- The relation of type $is–a$, which is the strictest; it is established in the case where each frame element (FE) is linked to a corresponding element of a subordinate frame.

- The $Using$ relation: indicates the case where the subordinate frame uses the parental frame as context, for example, the SPEED frame evokes the MOVEMENT frame.

In this case, it is not obligatory that all the FEs of the parental framework are linked with the elements of the subordinate framework.

- The relation $Subframe$ writes the subordinate frame as a sub-event of a larger event, for example, for the CRIMINAL TRIAL frame, the subordinate frames are ARREST, COURT OF JUSTICE, JUDGMENT.

- The relation $Perspective$ on means that the subordinate frame qualifies the general, non-oriented point of view of the parent frame. For example, the HIRE and GET A JOB frames are sub-frames of the EMPLOYMENT START frame from the respective perspectives of the employer and worker.

The other relationships used in FrameNet are anticipation and causation. Let us note that these relations are carried out in double voice, active and passive, where the roles of the actors change places. There are currently extensions to the FrameNet project in seven languages, but not yet in French.

3.3.7.2 WordNet

The English WordNet thesaurus (242, 243, 244) showed up on the Internet in 1995 yet its advancement was started at Princeton University 1984 under the way of psycholinguist George Miller. Its form 3.0 incorporates around 155,000 lexemes with models coordinated in 117,000 arrangements of equivalents called synsets, for the English language. Every synset can be viewed as the lexicalized show of an overall thought, or idea.

WordNet is built from the synonymy relation. For the designers of the thesaurus, two expressions are synonymous if the replacement of one by the other does not change the truth value of the proposition. However, the substitutability of synonyms in all contexts is not necessary: it is sufficient that the synonyms are replaceable in certain contexts. This makes it possible to admit that the same lexeme can be associated with several synsets, which corresponds to the flexibility of natural languages. There are currently over 200,000 lexeme-meaning pairs in WordNet. From WordNet version 2.0, we introduced the relationships between synsets that have the same root, i.e. semantically linked, but which belong to different parts of speech. This option was introduced to make the WordNet model more universal and less dependent on the specifics of different languages.

The definition of synonymy by substitutability required the division of WordNet according to the parts of speech: nouns, adjectives, verbs, adverbs. The descriptive structure of each part of speech is different from the others. Nouns are organized in hierarchical systems where properties from higher levels are inherited from lower levels. The hierarchy of names is embodied in the form of relationships of three types: hyperonymy-hyponymy (ie is-a), anonymity, and meronymy-holonymy (ie part-of).

In WordNet, verbs are distinguished from each other according to the context of the semantic field. There are three general categories of verbs: verbs denoting actions, verbs denoting events, and verbs denoting states. The first group of verbs is divided by 14 semantic fields among which there are the verbs of movement, change, possession, etc. three types of relationships are established between verbs in WordNet: implication, toponymy[1], and the causal relation. Nevertheless, the actors recognize that there is no strict delineation between the classes of verbs.

[1]The troponimic relation corresponds to the pattern: "verb-2 describes more precisely the action of verb1"

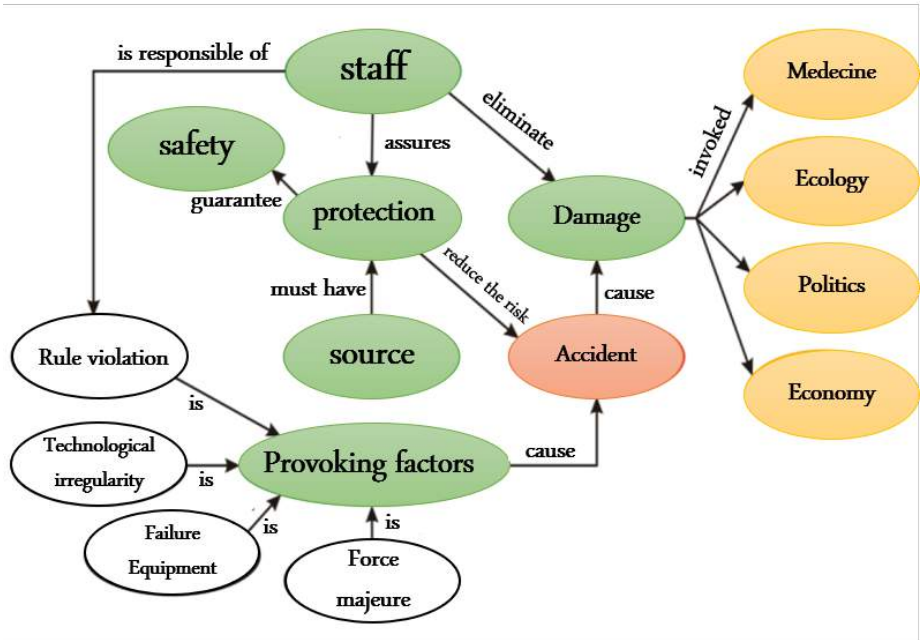


Figure 3.4: Fragment of the domain model

3.3.7.3 PyTe3 (RuTez)

The Russian PyTe3 thesaurus has been constructed since 1994 at the Information Processing Research Center of Lomonosov University in Moscow (245). Currently, it includes more than 51,500 concepts (general notions), more than 155 million lexical entries (words or sentences), and more than 200,000 relations between concepts. In total, taking into account the hierarchy of links, the thesaurus includes more than 2 million relations between concepts. The concepts of thesaurus also have lexical entries in English with over 125,000 words and phrases.

From the start, PyTe3 was designed to automate search engine searches for information, including solving the problem of ambiguity in user queries. The principle of its operation is formulated as follows: among the potential relations of a concept, one can rely on the relations which, in all observations of the entities of the concept (or in the great majority of them), do not disappear and do not change. For example, every forest is made up of trees.

This thesaurus uses several types of relationships. The first is substitution having the properties of transitivity and succession. The second is meronymy – holonymy; it applies to the components of an object, and also to the description of its intrinsic properties, and to the role that the object plays in a given situation. An important constraint is that each meronym always respects this relation with its concept-holonomy, and not with the other concepts. This ensures the transitivity of the relationship. For example, if it is true that a branch is part of a tree and a tree is part of a forest, we cannot say that a branch is part of

a forest.

Another type of relationship in the PyTe3 thesaurus is what the authors call asymmetric association. This is the case where a concept would not exist without another concept. For example, the concept of "summit of state" requires the existence of the concept of "head of state". The last type of relationship is the symmetrical association that links concepts that are very close but that the authors have not dared to bring together in the same concept. PyTe3 is designed for applications in the social and political fields. But, according to the authors, it ensures good accuracy of results in information retrieval for a wide range of more general topics.

3.3.7.4 BabelNet

The designers define BabelNet as an "encyclopedic dictionary" providing lexicalized concepts and named entities that are linked by various semantic relationships (246) and this in several languages. BabelNet encodes information as a coordinated diagram marked $G = (V, E)$ where V is the arrangement of hubs, each comparing to an idea or a named element and $E \subseteq V \times R \times V$ is the arrangement of curves connecting idea sets. Each arc has a semantic relation that can be specified in WordNet, such as, for example, is-a, part-of, or unspecified. Each node $v \subseteq V$ contains the lexicalization of the concept in the different languages. The authors call these multilingual concepts Babel synsets. BabelNet's particular interest in ontology learning is that this resource is built automatically by aligning WordNet's synsets with Wikipedia pages that act as a disambiguated context.

3.4 Ontology learning " layer cake "

The construction of an ontology involves several steps; a whole set of techniques is used to acquire the elements constituting the ontology. One of the tasks is the integration of the isolated terms into a hierarchical structure of concepts, provided with rules of inference and axioms, traditionally a strategy is used as under the name of "layer cake" (247); several stages follow one another, and at each stage, the results of the previous stage are used as inputs.

Examples:

- axioms and rules: $\forall x, y(absorbedBy(x, y) \rightarrow radioactive(x))$
- non-taxonomic relationship $cure(domain : Cancer, range : Radiation)$
- taxonomic relationship $is - a(Radionuclide, Element)$
- concepts $Element := \langle I, E, L \rangle$

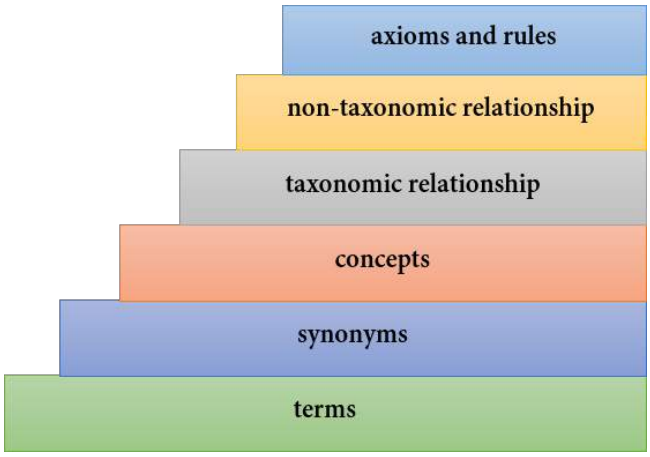


Figure 3.5: Ontology learning ”layer cake”

- synonyms {*Irradiation*, *exposure*}
- terms: {*Irradiation*, *accident*, *exposure*}

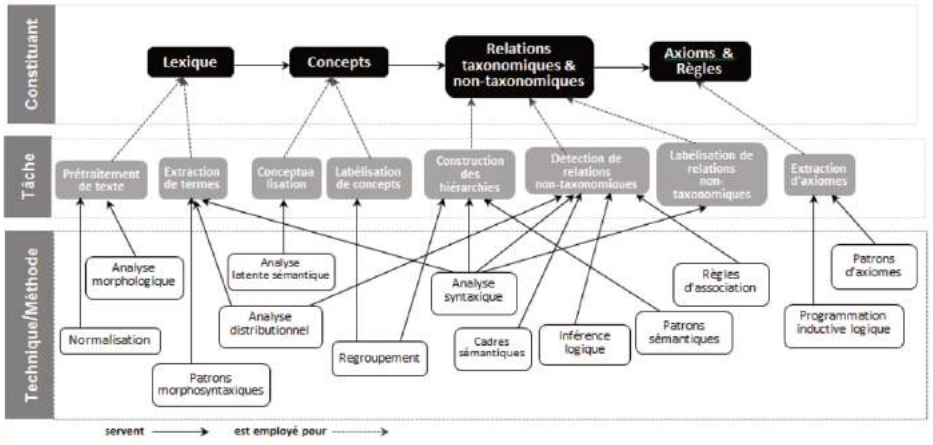


Figure 3.6: Sequence of tasks, indication of techniques adopted for each and ontology elements produced

This Layer Cake strategy has been implemented more fully in the Text2Onto platform. The Layer Cake approach has been criticized by authors in (248). The authors emphasize in particular that as the ontology construction stages take place one after the other in a pre-defined order, the conceptualization becomes completely dependent on the preceding result. Nevertheless, according to authors in (247), Layer Cake makes it possible to discover the important notions of the domain, and their names remain reasonable.

But, for us, preliminary work with experts in the field is necessary in order to start the work from the most important concepts; text mining and linguistic analysis are called upon to facilitate this work.

3.4.1 Terminology, a specialized sub-language

Ontology engineering is particularly interested in terminology since the domain ontology is built from the terms and, more broadly, from the domain lexicon. Terminology aims to provide a common standard for terminology work. It is a discipline using several theories and several methods. Its recommendations are taken into account by the International Organization for Standardization (ISO), and in particular by ISO Technical Committee (which gets ready norms and different archives concerning the system and working standards on wording and assets. etymological). The origins of these standards come from the work of Eugene Wuster and the Vienna terminology school he facilitated (249). In (250) the authors distinguish three categories of tasks that are specific to the field of computational terminology:

1. Automatic identification and filtering of terminology candidates.
2. Grouping of variations and synonyms.
3. Tracing the relationships that bind them.

These problems are therefore close to those that we aim to solve for the learning of ontologies. The terminology of a domain is considered as a sub-language inscribed in natural language and characterized by two aspects (249, 251, 252):

1. The cognitive aspect that links the linguistic forms (the terms themselves) and their theoretical substance, for example the referents in reality.
2. The linguistic aspect which gives the syntactic rules to which the terms are subject. The methods that we are going to discuss relate to these two dimensions. Whatever their targets (terms, concepts, relationships), learning strategies can be partitioned into three classes: measurements, etymology, and half and halves (253). But all learning methods use probabilistic models based on statistical measures. Therefore, when we say ”linguistic methods” we want to emphasize that additional linguistic information is taken into account. For example, all words are ”labeled with tags indicating their part of speech, that is, their syntactic function in the sentence. We will qualify as ”statistical methods” that rely solely on raw texts to discover the searched entities.

3.4.2 Extraction of terms

In this section, we present the methods of extracting terms from the textual corpus, whether or not they use additional resources. Terms can be made up of single words, or sequences of words. In technical fields, most terms are compound (or multi-word) terms. For example, in (254), the authors have shown that the terms of Radiological safety comprise on average

3 lexical units. In linguistics, terms are hierarchical syntactic combinations of words, called terminological phrases, or synapsies (255). This means that, for the acquisition of terms, the methods must be more 'elaborate than those used for other Data Mining tasks such as the extraction of themes (Topic Extraction) or all the lexical features, words, lemmas, or stemmes, can be considered as components of a thematic (256). Note another difference between the methods of extracting terms from the textual corpus and those of DataMining: if we want to be able to extract compound terms, we must not use certain traditional steps of preprocessing texts such as cutting endings (stemming) or the elimination of “stop words”.

3.4.2.1 Frequency-based methods

In these methods, we assume that the frequencies of occurrence of the domain terms are particularly high. In the following, we cite the main characteristics that are used for the selection of candidate terms. Table 3.2 summarizes their calculation formulas. To describe them we will utilize the accompanying documentations:

- $TF(w)$ is the recurrence of a word or an expression in the corpus.
- $DF(w)$ is the quantity of records that contain the word or expression w in some measure once.
- $TF(w|d)$ is the recurrence of a word or expression w in record d.
- $|W|$ is the complete number of words or expressions in the corpus.
- $|D|$ is the absolute number of archives in the corpus.
- $|W_d|$ is various words or expressions in archive d.

The first two characteristics are the term frequency (TF) and the presence/absence of a term in a document (Document frequency (DF)) or in all the documents of the corpus. These measurements are used for the calculation of more complex characteristics.

- **Term Frequency Inversed Document Frequency (TF-IDF)** favors words or phrases that are often encountered in a limited subset of documents in the corpus.
- **Term Frequency Residual Inverse Document Frequency (TF-RIDF)** is a modification of the previous measure which uses Poisson's law by assuming that the distribution of search terms, unlike common words, does not correspond to that of Poisson (257).
- **Domain Consensus (DC)** is based on the entropy calculation; this characteristic favors the more frequent words or phrases in certain documents of the corpus (258).

	Formula
TF-IDF	$TF(w) \times \log \frac{\|D\|}{\|DF(w)\|}$
TF-RIDF	$TF(w) \times (\log \frac{\|D\|}{\|DF(w)\|} - (-\log(1 - e^{-\frac{\|TF(w)\|}{\|D\|}})))$
Domain Consensus	$-\sum_{d \in D}(\frac{TF(w\|d)}{\|Wd\|}) \times \log \frac{TF(w\|d)}{\|Wd\|})$
Term Contribution	$\sum_{d_i, d_j \in D : d_i \neq d_j} (TF(w\|d_i) \times log \frac{\|D\|}{DF(w)}) \times (TF(w\|d_j) \times log \frac{\|D\|}{DF(w)})$
Term Variance	$\sum_{d \in D} (TF(w\|d) - \frac{TF(w)}{\|D\|})$
Term Variance Quality	$\sum_{d \in D} TF(w\|d)^2 - \frac{1}{\|D\|}(\sum_{d \in D} TF(w\|d))^2$

Table 3.2: Summary of frequency characteristics for term extraction

- **Term Contribution (TC)** on the contrary penalizes frequent words which are uniformly distributed in the documents of the corpus.
- **Term Variance Quality (TVQ)** penalizes words and phrases that occur only once in most of reports in the corpus; TVQ corresponds to the variance of the frequency of the term among the documents where the word (or the phrase) is encountered at least once (259). Words or phrases which are present in few corpus documents, or which are uniformly distributed among these documents, have low values for Term Variance.

3.4.2.2 Methods based on contrast corpora

These methods are based on the comparison of the frequency distribution of words in the specialized corpus (or targeted corpus) and the reference corpus (or contrasted corpus) which contains the more general texts: it is assumed that the behavior of the terms in both collections is significantly different. For the results obtained to be correct, the corpora must have comparable sizes. To measure these characteristics, we complete the previous list:

- $TFr(w)$ is the frequency of the word or phrase w in the reference corpus.
- $DFr(w)$ is the quantity of reports in the reference corpus (s) which contain the word or expression w.
- $|Dr|$ is the total number of documents in the reference corpus (s).
- $|Wr|$ is the all out number of words or expressions in the reference corpus (s).
- $|Cw|$ is the quantity of corpora that contain the word or expression w.

- $|C|$ is the quantity of corpora to analyze including differentiating corpora.

The basic characteristic in this group of measures is Weirdness, which is the ratio between the relative frequencies of the word w in the target corpus and the contrasting corpus (260).

$$weirdness(w) = \frac{TF(w)}{|W|} / \frac{TF_r(w)}{|W|_r} \tag{3.1}$$

The **Relevance** index exploits the same principle: infrequent words and phrases have low values for this characteristic. Conversely, Relevance has a significantly high value if the expression is frequent, but not too frequent, in the targeted corpus nor too rare in the texts of the contrasted corpus (261).

$$Relevance(w) = 1 - \frac{1}{\log_2(2 + \frac{TF(w) \times DF(w)}{TF_r(w)})} \tag{3.2}$$

The **TF-IDF** measure 3.2 is part of the methods of this group but, here, we calculate IDF on the contrasted corpus. Its new expression is **KF-IDF** which favors words and phrases more frequently in the targeted corpus than in the contrasted corpus (262).

$$KF - IDF(w) = DF(w) \times \log(\frac{|C|}{|C_w|} + 1) \tag{3.3}$$

Two other modifications of TF-IDF are **Contrastive Weight** (263) and **Discriminative Weight** (264). The first concerns the hypothesis that the distribution of common words must be the same in the two corpora.

$$CW(w) = \log TF(w) \times \log(\frac{|W|+|W_r|}{TF(w) + TF_r}) \tag{3.4}$$

Discriminative Weight penalizes frequent words in the targeted corpus if they are more specific in the contrasted corpus.

$$DW(w) = DP(w) \times DT(w) \tag{3.5}$$

or

$$DP(w) = \log_{10}(TF(w) + 10) \times \log_{10}(\frac{|W|+|W_r|)}{TF(w) + TF_r(w)}) \tag{3.6}$$

and

$$DT(w) = \log_2(\frac{TF(w) + 1}{TF_r(w) + 1} + 1) \tag{3.7}$$

The number **Loglikelihood** favors isolated words, and groups of words, the relative frequency of which is higher in the targeted corpus than in the contrasted corpus (265).

$$loglikelihood(w) = 2 \times (TF(w) \times \log \frac{TF(w)}{TF^{\exp}(w)} + TF_r(w) \times \log \frac{TF_r(w)}{TF^{\exp}(w)}) \tag{3.8}$$

or

$$TF^{\exp}(w) = |W| \times \frac{TF(w) + TF_r(w)}{|W| + |W_r|} \tag{3.9}$$

and

$$TF_r^{\exp}(w) = |W_r| \times \frac{TF(w) + TF_r(w)}{|W| + |W_r|} \tag{3.10}$$

3.4.2.3 Methods based on the measurement of association between words

These traditional methods use measures of association between two words, or between two fragments of text, to estimate the mutual correlation of two candidates in terms. They are intended for extracting terms consisting of more than two words; they are not applicable for the extraction of terms consisting of isolated words. Note that the association measures can also be used for conceptualization, as well as for the extraction of relations between concepts. In the formulas in this section, we will use the following notations:

- $\bar{x}$ is any word other than word x.
- $r(w)$ and $l(w)$ is the all out number of various words that are previously (left) and after (right) of the word w.

Mutual Information, provided by (266). The continues to be widely used for term extraction because it is simple and clear.

$$MI(w, y) = |W| \times \log \frac{TF(x, y)}{TF(x) \times TF(y)} \tag{3.11}$$

If the words x and y are encountered in the corpus independently of each other, their mutual information is zero. On the contrary, the higher the value of MI, the more these words are semantically related. The above formula is not suitable for the case where one of the words is missing. Several works have proposed modifications to equation 3.11 in an attempt to solve this problem. For example, in (267), the authors propose the so-called "enhanced" measure, **Enhanced Mutual Information** defined as the ratio between the probability of occurrence of the pair of words and the product of the probabilities of the individual occurrences of every word outside of the pair.

$$EMI(w, y) = \log \frac{TF(x, y)}{(TF(x) - TF(x, y))(TF(y) - TF(x, y))} \tag{3.12}$$

The measurement of normalized mutual information **Normalized MI** has been proposed in (268); it improves results in the case of low frequencies of occurrence of words:

$$NormalizedMI(x, y) = \frac{\log \frac{|W| \times TF(x,y)}{TF(x) \times TF(y)}}{-\log \frac{TF(x,y)}{|W|}} \tag{3.13}$$

Cubic MI is another modification of MI; it was proposed in (269). The same goal, that of adapting the formula to the case of infrequent co-occurrences

$$CubicMI(x,y) = \log \frac{TF(x,y)^3}{|W| \times TF(x) \times TF(y)} \tag{3.14}$$

The author in cite daille1994approche also proposed the textbf True MI measure:

$$TrueMI(x,y) = TF(x,y) \times \log \frac{TF(x,y)}{TF(x) \times TF(y)} \tag{3.15}$$

In (270), the **Symmetrical Conditional Probability** allows to check the cohesion between the words x and y:

$$SCP(x,y) = \frac{TF(x,y)^2}{TF(x) \times TF(y)} \tag{3.16}$$

In (271), the authors propose to use the Dice coefficient **Dice coefficient**, initially borrowed from information theory for machine translation.

$$DC(x,y) = \frac{2 \times TF(x,y)}{TF(x) + TF(y)} \tag{3.17}$$

In (272), the modification proposed by authors improved the quality of extraction of frequent bigrams, and their classification.

$$ModifiedDC(x,y) = \log(TF(x,y)) \times \frac{2 \times TF(x,y)}{TF(x) + TF(y)} \tag{3.18}$$

A new modification and generalization of **Mutual Information** has been proposed in (273). The authors propose a measure to calculate the cohesion of terms composed of several words ($n > 2$) and assign the highest values to terms with the highest frequency of co-occurrence.

$$GeneralizedDC(w,y) = \frac{r_w \times \log TF(w) \times TF(w)}{TF(y)} \tag{3.19}$$

In (269), the author also adopted several measures of association, borrowed from information theory, for the extraction of two-word terms; among them the **Simple Matching Coefficient** (SMC), the **Kulczinsky coefficient** and the **Yule coefficient**. The **SMC** adds the numbers of co-occurrences, and disjoint occurrences, of two words (independently of each other).

$$SMC(x,y) = \frac{TF(x,y) + TF(\bar{x},\bar{y})}{|W|} \tag{3.20}$$

The **Kulczinsky coefficient** varies from 0 to 1 and, in the case where the word x only appears with the word x, it is greater than 0.5:

$$KulczinskyCoefficient(x,y) = \frac{TF(x,y)}{2}\left(\frac{1}{TF(x)} + \frac{1}{TF(y)}\right) \tag{3.21}$$

The **Yule coefficient** varies from -1 to +1. If the words are independent in the corpus, it is equal to zero. For words that are always together its value is equal to +1 and for words that are never together, it is equal to -1.

$$YUC(x,y) = \frac{TF(x,y) \times TF(\bar{x},y) - TF(x,\bar{y}) \times TF(\bar{x},y)}{TF(x,y) \times TF(\bar{x},y) - TF(x,\bar{y}) \times TF(\bar{x},y)} \tag{3.22}$$

To extricate multi-word terms, we can likewise utilize the **Jaccard coefficient** which is the proportion of the quantity of co-events of two words and the amount of the events of every one of them without the other.

$$JaccardCoefficient(x,y) = \frac{TF(x,y)}{TF(x,\bar{y}) + TF(\bar{x},y)} \tag{3.23}$$

The measurements of **Chi-Square** and **T-Score** are used to measure the degree of dependence of two words composing a phrase.

$$Chi - Square(x,y) = \frac{\left(TF(x,y) - \frac{TF(x)\times TF(y)}{|W|}\right)^2}{TF(x) \times TF(y)} \tag{3.24}$$

$$T - Score(x,y) = \frac{TF(x,y) - \frac{TF(x)\times TF(y)}{|W|}}{\sqrt{TF(x,y)}} \tag{3.25}$$

The **Gravity Count** proposed in (274) is a measure of associativity which makes it possible to estimate the frequency at which the second word in the pair follows (i.e., appears on the right) the first word or the reverse.

$$GravityCount(x,y) = \log\frac{TF(x,y)r(x)}{TF(x)} + \log\frac{TF(x,y)l(y)}{TF(y)} \tag{3.26}$$

The last parametric test tool we will cite for term selection is **LogLikelihood Ratio**. It compares the maximum values of two likelihood functions according to the assumptions that the two words form or not a phrase (275).

$$\begin{aligned} LLR(xy) = 2 \times (&TF(xy) \times \log\frac{|W|\times TF(xy)}{TF(x) \times TF(y)} + TF(x\bar{y}) \times \log\frac{|W|\times TF(x\bar{y})}{TF(x) \times TF(\bar{y})} \\ &+ TF(\bar{x}y) \times \log\frac{|W|\times TF(\bar{x}y)}{TF(\bar{x}) \times TF(y)} + TF(\bar{x}y) \times \log\frac{|W|\times TF(\bar{x}\bar{y})}{TF(\bar{x}) \times TF(\bar{y})}) \end{aligned} \tag{3.27}$$

3.4.2.4 Context-based methods

As their name indicated, those methods are context-based for locating terms. Here, the context is defined by isolated words or fragments composed of several words which serve as limits between which the terms are regularly found. With this approach, we speak of “nested terms” In practice, the values of the bounds are heuristic; they can be domain-specific or more general. Below, we will list the measures for choosing the candidate selection threshold in terms identified in the text using the context. The notations used in the formulas are:

- P_w is the arrangement of all sentences that contain the word or expression w.
- C_w is the arrangement of all settings of the word or expression w.
- $|W_c|$ is the quantity of context oriented words in the word or expression w
- $TF_w(c)$ is the recurrence of the word c as a setting expression of the word or expression w.
- $F_{max}(w)$ is the most extreme worth of the recurrence of a n-gram that contains the word or expression w (for example a setting in addition to the word or expression w).
- P_w^N is the set N of the most regular n-grams that contain the word or expression w.
- $l_{token}(w)$ and $r_{token}(w)$ are the amounts of the frequencies of the words which, in the texts, are to the left, individually to the right, of the word or of the expression w.
- $ltype(w)$ and $rtype(w)$ is the quantity of one of a kind logical words that are simply to the left, individually to the right, of the word or expression w.
- $|w|$ is the quantity of words in the expression w.

A model exploiting the context was presented in (276). The characteristic used in the methods of this group is **C-Value**. Initially, it was used to find terms consisting of several words. Thus*C-Value* favors longer-term candidates and penalizes phrases that frequently fit inside noun groups.

$$C-Value(w) = \begin{cases} log_2|w| \times \left(TF(w) - \frac{\sum_{p \in P_w} TF(p)}{p_w} \right) & , if\ the\ sentence\ envelops\ w \\ log_2|w| \times TF(w) & , otherwise \end{cases} \tag{3.28}$$

In (277), we find a generalization of this measure for single-word terms:

$$C - Value(w) = TF(w) - \frac{\sum_{p \in P_w} TF(p)}{|P_w|} \tag{3.29}$$

The most well-known modification of **C-Value** is **NC-Value** which adds contextual information to (C-Value); the following equation makes it possible to estimate the rate of independence of each word in the text or, in other words, to check whether the word m is necessarily associated with the other words (277).

$$C - Value(w) = \frac{1}{|W|} \times MC - Value(w) \times cweight(w) \tag{3.30}$$

$$cweight(w) = \sum_{e \in C_w} weight(c) + 1 \tag{3.31}$$

$$weigh(c) = \frac{1}{2} \left(\frac{|W_c|}{|W|} + \frac{\sum_{e \in W_c} TF(e)}{TF(c)} \right) \tag{3.32}$$

In (278), the authors propose another variation of **NC-Value**.

$$NC - Value(w) = 0.8 \times C - Value(w) + 0.2 \times \sum_{c \in C_w} TF(c) \tag{3.33}$$

The **Insideness** and **SumN** (279) have measures that consider words and phrases in the context of the sentences (sentences) that surround them. Other measures consider words and phrases in the context of the sentences around them, for example **Insideness** and**SumN** (279).

$$Insideness(w) = \frac{TF_{max}(w)}{TF(w)} \tag{3.34}$$

$$SumN(w) = \frac{\sum_{p \in P_w^N} TF(p)}{N \times TF(w)} \tag{3.35}$$

The **Insideness** allows finding the parts (fragments) of real terms, while **SumN** allows checking if a word or a phrase is useful for the construction of the domain lexicon. Here are yet other measures built on the assumption that certain words are used more often in units of terminology. We measure the increase in the probability that the fragments containing these units are themselves terms. The authors in (280) propose the two formulas: **Token-LR** and **Type-LR**.

$$Token - LR(w) = \sqrt{l_{token}(w) \times r_{token}(w)} \tag{3.36}$$

$$Type - LR(w) = \sqrt{l_{type} \times r_{type}} \tag{3.37}$$

But these two measures only consider contextual words, without taking into account the candidate terms themselves. In the same book, in order to overcome this drawback, the FLR measurement was proposed, which also has two variants: **Token-FLR** and **Type-FLR**.

$$Token - FLR(w) = TF(w) \times Token - LR(w) \tag{3.38}$$

$$Type - FLR(w) = TF(w) \times Type - LR(w) \tag{3.39}$$

3.4.3 Synonyms and Multilingual Variants

The equivalent word level tends to the obtaining of semantic term variations in and between dialects, where the last option indeed concerns the procurement of term interpretations. A large part of the work in this space has zeroed in on the coordination of WordNet[1] for the securing of English equivalents, and EuroWordNet[2] for bilingual and multilingual equivalent words and term interpretations. A significant part of this work is the ID of the suitable (WordNet/EuroWordNet) feeling of the term being referred to, which decides the arrangement of equivalents that are to be separated. Clearly, this includes standard word sense disambiguation calculations, the vast majority of which depend on (281) assessment campaigns[3] for late methodologies on word sense disambiguation). Nonetheless, explicitly in the cosmology learning setting, specialists have taken advantage of the way that equivocal terms have quite certain implications specifically areas taking into consideration a coordinated way to deal with sense disambiguation and space explicit equivalent word extraction (look at (247, 282, 283, 284, 285).

As opposed to utilizing promptly accessible equivalent sets, for example, given by WordNet and related lexical assets, specialists have additionally dealt with calculations for the powerful securing of equivalent words by grouping and related methods. On this premise, much work has been done on equivalent word securing from text corpora that depends on Harris' distributional theory that terms are comparable in significance to the degree in which they share syntactic settings (286) and Reinberger and Spyns (this volume). Related work starts out of term ordering for data recovery, for example the group of Latent Semantic Indexing calculations (LSI). LSI and related methodologies apply aspect decrease procedures, for example, those depicted in (287) to uncover innate associations between words, consequently

[1]WordNet is unreservedly open from hap/wordnet.princeton.edu

[2]EuroWordNet can be authorized from ELAM at http://www.elda.fr

[3](see additionally the SEN-SEVAL (http://www.senseval organization/)

prompting bunch arrangement. Truth be told, LSI-based methods are particularly fascinating as they don't run into information meager condition issues, for example, approaches depending on crude information. At last, given the developing significance of the web in information securing, there is by all accounts a latest thing to utilize factual data measures characterized over the web to recognize equivalents, models are given in (288) and in (289).

3.4.4 Concepts

The extraction of ideas from text is questionable as it isn't clear what precisely establishes an idea. In our view, idea enlistment or development ought to give:

- A purposeful meaning of the idea.
- A bunch of idea cases, for example its augmentation.
- A bunch of etymological acknowledge, for example (multilingual) terms for this idea.

In this manner, we characterize an idea as a couple with dictionary $(\Psi, \Sigma) \oplus L$ where Ψ is the intension of the idea, Σ its augmentation, and L portrays its phonetic acknowledgment. The greater part of the exploration in idea extraction resolved the inquiry from an etymological or text based viewpoint, seeing ideas as groups of related terms. Clearly, this methodology covers totally with that of term and equivalent word extraction as talked about above.

On the other hand, scientists have resolved the issue according to an extensional perspective, a model is given in (290) determined chains of command of named substances from text and in this way finding ideas according to an extensional perspective. The Know-It-All framework (291) likewise targets learning the expansion of ideas such as all film entertainers showing up on the Web. In the methodology of (290) the ideas, just as the augmentation, are inferred at the same time, while (291) basically populates existing ideas with occurrences. Note that in this regard, the philosophy populace is a lot of identified with cosmology learning.

At last. deliberate idea learning incorporates the extraction or procurement of formal and casual definitions. A casual definition may be a printed portrayal, for example a shine of the idea. A conventional definition incorporates the extraction of idea properties. some portion of which is the extraction of relations between a specific idea and different ideas. The extraction of casual idea definitions is very uncommon. Indeed, the main work detailed in this space is the OntoLearn framework that determines WordNet-like shines for area explicit ideas. The extraction of formal idea definitions. taking everything into account will be examined in the following two segments.

3.4.5 Taxonomy

There are as of now three primary ideal models took advantage of to initiate scientific classifications from text based information. The first is the use of lexico-syntactic examples to distinguish hyponymy relations as proposed by (292). Nonetheless, it is notable that these examples happen seldom in corpora. In this manner, however approaches depending on lexico-syntactic examples have a sensible accuracy, their review is exceptionally low. Identified with this are likewise moves toward that exploit the inner construction of thing expressions to infer ordered relations between classes communicated by the top of the thing expression and subclasses can be gotten from a blend of the head and its modifiers (293). The subsequent worldview is again founded on Harris' distributional speculation, as talked about above with regards to equivalent word extraction and term bunching. In this line, individuals have predominantly taken advantage of progressive bunching calculations to consequently get term pecking orders from text, a model is given in (294). The third worldview originates from the data recovery local area and depends on a report based thought of term subsumption as proposed for instance in (295).

3.4.5.1 Extracting relations - a multi-level task

According to authors in (296), the definition of the relationships between the elements of ontology can be considered from several angles, linguistic, semantic, and ontological, each contributing to the interpretation of the constituents of ontology and their relationships. The last step in ontological learning is the implementation of the results in a formal language. At the semantic level, the relationships between concepts as well as the assignment of properties to concepts are presented through predicates. This fact implies the need, beforehand, to list, classify and structure them.

3.4.5.2 Classification of relations

For learning ontologies, relationships can be classified according to two aspects: semantics and ontological. According to the formal definition (297), the ontology can be described as follows: $O = (C, H, I, R, P, A)$ In this 6 – tuplet are listed the following elements: $C = C_C \cup C_I$ is the set of classes, or concepts (C_C) each of which corresponds to the set of entities (C_i) that populate this concept. H, I, and R are the types of relations; P is the set of properties of concepts and A is the set of axioms and rules which control the lucidness of philosophy and permit the surmising of new information.

$$H = kind\_of(c_1, c_2) | c_1 \in C_c, c_2 \in C_c \tag{3.40}$$

$$I = is\_a(c_1, c_2)|c_1 \in C_c \wedge c_2 \in C_c \tag{3.41}$$

$$R = rel\_k(c_1, c_2, ...c_n)|\forall i, c_i \in C \tag{3.42}$$

$$P = prop\_(c_k, datatype)|c_k \in C_c \wedge prop\_I(c_k, value)|c_k \in C_I \tag{3.43}$$

$$A = condition_x\ conclution_y(c_1, c_2, ...c_n)|\forall j, c_j \in C_c \tag{3.44}$$

We distinguish the two types of relations between concepts: hierarchical relations of subordination (or subsumption) and associative relations (non-taxonomic, non-subordinate). Hierarchical relationships, themselves, fall into two subcategories: Kind-of or Class – Subclass, and Is-a or Class – Entity.

When implemented in a formal language, the relationships that have been highlighted are assigned certain properties, such as transitivity, symmetry, order, equivalence; at the same time, restrictions are imposed on them. This set of ontological relationship types is not sufficient to satisfactorily model a specialized field, nor to provide rules for detecting linguistic features of relationships in texts.

However, to our knowledge, there is currently no universal classification of predictive relationships, although some relationships are unanimously accepted: synonymy, antonymy, causality, hierarchical relationships such as hyperonymy, meronymy, etc. One of the first systematizations of terminology as a formal sub-language was carried out by authors in (298) who proposed to distinguish between logical and ontological relations. In each group of relationships, there is a hierarchy. Logical relations are defined as a relation of resemblance. Ontological relations are understood as relations of contiguity in space and time.

3.4.6 Rules

The extraction of rules is most likely the most un-tended to investigated region in cosmology learning. Introductory outlines for this assignment can be found for instance in PSI. Further, the new PASCAL lexical entailment challenge[1] (299) addresses a connected issue. truth be told, this test has unequivocally expanded the attention to the issue of inferring lexical entailment rules and lead numerous analysts to resolve the issue, so a plenty of ways to deal with tackle the issue of taking in ontological principles from text corpora can be anticipated soon. The principle concentrate thusly has been to learn lexical entailments for application being referred to noting frameworks, see (286).

[1] httpl/www.pascal.networking/Challenges/RTE/

Category of relations	Group of relations	Relation	Type of concept
Qualitative	Hierarchy	Kind-of (Is-a)	Abstract and concrete
		Attribute ↔ Attribute value	
		Invariant ↔ Variant	
	Aggregation	Entire ↔ Part (Part-of)	Membership
		Object ↔ Property (s)	
		Object ↔ Location	
		Level ↔ level unit	
	Functional	Action object ↔ Action ↔ Action subject	Process
		Cause ↔ Consequence	
		Condition ↔ Action	
		event ↔ Action	
		Aspect (E) ↔ Action	
		event ↔ Aspect (State)	
		Term ↔ Synonym	
		Data ↔ Action	
		Data ↔ Quantities	
	Semiotics	Term ↔ Mode of use	The background and the form
		Term ↔ Mode of representation	
		Term ↔ Term sign	
Quantitative	Equivalence	Term ↔ Term synonym	Equivalence and opposition
	Correlation	Term ↔ Correlative of term	

Table 3.3: Types of relationship between terms.

3.5 Ontology Engineering Tools and Environments

3.5.1 Ontolingua Server

Ontolingua Server (300) is an electronic help that is expected to give a typical stage wherein ontologies created by various gatherings can be shared, and maybe a typical perspective on these ontologies accomplished. The focal part of Ontolingua is a library of ontologies, communicated in the Ontolingua metaphysics definition language (in light of KIF). A server program gives admittance to this library. The library might be gotten to through the server in more than one way: either by altering it straightforwardly (by means of an electronic interface) or by utilizing programs that contact the server remotely by means of a NGFP interface. The Ontolingua server was prepared to do naturally changing ontologies communicated in one arrangement into an assortment of others (e.g., the CORBA Interface Definition Language-IDL).

3.5.2 OntoEdit

OntoEdit (301) Collaborative metaphysics advancement for the semantic web Ontologies currently assume a significant part in empowering the semantic web. They give a wellspring of exactly characterized terms for example for information serious applications. The terms are utilized for brief correspondence across individuals and applications. Normally the improvement of ontologies includes synergistic endeavors of various people. OntoEdit is a cosmology manager that coordinates various parts of metaphysics designing. This paper centers around the shared improvement of ontologies with OntoEdit which is directed by a far reaching strategy.

3.5.3 Protégé

Protégé (302) is an extensible, stage free climate for making and altering ontologies and information bases created by the Stanford Medical Informatics (SMI) at Stanford University. It is an open-source, independent application with an extensible engineering. The center of this climate is the philosophy supervisor, and it holds a library of modules that can be stopped, called modules, to add more capacities to the climate. The principle Protégé capacities are to load and save OWL and RDF ontologies; alter and imagine classes, properties, and SWRL rules; characterize legitimate class attributes as OWL articulations; execute reasoners, for example, depiction rationale classifiers, and alter OWL people for Semantic Web markup. Protégé is accessible in various forms, each including distinctive modules, whose primary contrast is the philosophy language that they support:

- Protégé version3 supports OWL1.0, RDF(S) and Frames.

- Protégé versions4 and 5 supports OWL2.0.

Protégé (303) is an open-source metaphysics editorial manager disseminated by Stanford University of Medical Informatics. It makes it conceivable to fabricate a metaphysics for a given area, to characterize information passage structures, and to gain information utilizing these structures as occurrences of this philosophy. Protégé is additionally an extensible stage, because of the arrangement of modules, which makes it conceivable to oversee sight and sound substance, question, assess and combine ontologies, and so on The Protected instrument has a graphical UI (GUI) permitting it to effectively control every one of the components of a metaphysics: class, property, occurrence, and so on Ensured can be utilized in any field where ideas can be displayed as a class order.

3.5.4 Neon Toolkit

The NeOn Toolkit (304) is the cosmology designing climate initially created as a feature of the NeOn Project and presently upheld, along with different advancements from NeOn, by the NeOn Foundation. The NeOn Toolkit is a best in class, open-source multi-stage philosophy designing climate, which offers exhaustive help for the cosmology designing life-cycle. The tool compartment depends on the Eclipse stage, a main improvement climate, and gives a broad arrangement of modules (presently 45 modules are accessible) covering an assortment of philosophy designing exercises, including Annotation and Documentation, Development, Human-Ontology Interaction, Knowledge Acquisition, Management, Modularization and Customization, Neon Plugins, Old Main Page, Ontology Dynamics, Ontology Evaluation, Ontology Matching, Reasoning and Inference, Reuse.

3.5.5 OntoUML Lightweight Editor (OLED)

The OntoUML lightweight manager (OLED) (300) is a climate for the turn of events, assessment, and execution of area ontologies utilizing the UFO-based ontologically very much established demonstrating language OntoUML. The apparatus gives a straightforward, lightweight, and coordinated arrangement of highlights to philosophy engineers, like linguistic confirmation, visual reenactment, model checking, model derivation, programmed semantic-enemies of examples location and amendment, approval of parthood relations and cosmology designs.

3.6 Ontology Languages and Formalisms

3.6.1 XML

Extensible Markup Language (XML) is a meta-language for addressing a text record in a tree structure utilizing a markup framework. XML thus provides a framework for structuring data

that can be employed for the rapid and syntactically unambiguous definition of a document format. This language was created to work with the trade, sharing, and distribution of information across the web. Hence, most of dialects/models presented for the semantic web are communicated in XML.

XML makes it conceivable to structure an archive by characterizing its own labels on a case by case basis and without considering either the importance of this construction or the PC frameworks that will work it. Principles like XPath and XQuery have been created to peruse and inquiry the XML tree design of records.

3.6.2 RDF

RDF is a standard model for information exchange on the Web. RDF has highlights that work with information consolidating regardless of whether the fundamental diagrams contrast, and it explicitly upholds the development of mappings after some time without requiring every one of the information buyers to be changed.

RDF broadens the connecting design of the Web to utilize URIs to name the connection between things just as the two closures of the connection (this is generally alluded to as a "triple"). Utilizing this basic model, it permits organized and semi-organized information to be blended, uncovered, and shared across various applications.

This connecting structure shapes a coordinated, named diagram, where the edges address the named interface between two assets, addressed by the chart hubs. This diagram view is the most straightforward mental model for RDF and is regularly utilized in straightforward visual clarifications.

(Asset Description Framework) is an information model for objects (assets) and the connections between them, giving basic semantics to this information model that can be addressed in XML.

RDF is, as well, a language for metadata on the web. The grammar of RDF depends on that of XML. The fundamental model of this language is intended to permit credits to be related with web assets utilizing semantic metadata portrayal. It is the detail of a framework for communicating straightforward semantic affirmations. Along these lines, RDF makes it conceivable to consider the Web to be an assortment of assets associated by joins named "semantically", and subsequently makes it conceivable to build the simplicity of programmed handling of Web assets.

The principal construction of any articulation in RDF is an assortment of triples, each comprised of a subject, a predicate, and an item.

- **Resource (Subject)**: a data substance that can be referred to by an identifier. This identifier should be a URI (Universal Resource Identifiers). The subjects or assets are

fluctuated: a page, some portion of a page, a total site, an item not open by the web like a book.

- **Property (predicate)**: a particular viewpoint, property, brand name or affiliation used to depict an asset.
- **Value**: A strict (single string) or an asset.

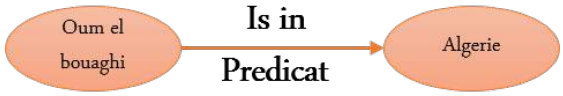


Figure 3.7: Triplet

A RDF record is a bunch of triples of the structure <subject, predicate, object>. The components of these triples can be URIs, literals, or factors. This set of triples can be represented by a graph. More exactly, a labeled oriented multigraph where the elements appearing as a subject where the objects are vertices, and each triplet is represented by an arc whose origin is its subject and the destination is its object. This document will be translated into an RDF / XML document, and is often represented in graphical form (see figure3.8).

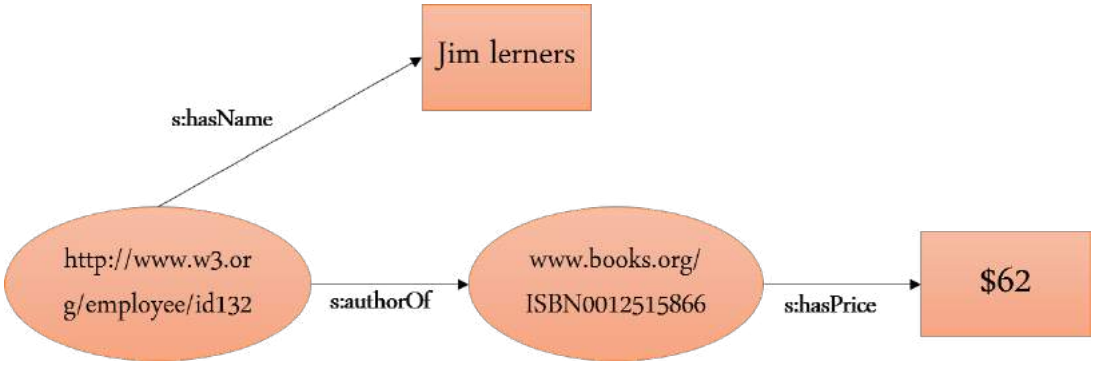


Figure 3.8: RDF graph example

RDFS is a W3C suggested meta-model for characterizing a pattern/model depicting the semantic universe of RDF articulations. RDFS is the jargon definition language for RDF, and in this way gives a composing framework to RDF explanations. It permits the meaning of classes and subclasses (rdfs: Class, rdfs: subClassOf) depicting the assets to be explained and giving significance to the properties related with the assets. It likewise permits the plan of limitations on the qualities related with a property to give it meaning (rdfs: area, rdfs:

range). Here is an example which defines a class ”student”, and which instantiates from this class a resource[1], and defines it as a subclass of class << *human* >>:

```
<rdfs:Class rdf:ID="student">
<rdf:type rdf:resource="http:www.ontoknow ledge.org/#DefinedClass"/>
<rdfs:subClassOf rdf:resource="#human"/>
</rdfs:Class>
```

3.6.3 RDF Schema

RDF Vocabulary Description Language 1.0: RDF Schema (RDFS): RDFS is a broadly useful language for addressing straightforward RDF vocabularies on the Web. Other jargon definition advancements, similar to OWL or SKOS, expand on RDFS and give language to characterizing organized, Web-based ontologies which empower more extravagant reconciliation and interoperability of information among graphic networks.

3.6.4 OWL

The W3C Web Ontology Language (OWL) is a Semantic Web language intended to address rich and complex information about things, gatherings of things, and relations between things. OWL is a computational rationale based language with the end goal that information communicated in OWL can be taken advantage of by PC programs, e.g., to confirm the consistency of that information or to make verifiable information unequivocal. OWL archives, known as ontologies, can be distributed in the World Wide Web and may allude to or be alluded from other OWL ontologies. OWL is important for the W3C’s Semantic Web innovation stack, which incorporates RDF, RDFS, SPARQL, and so on

The current rendition of OWL, additionally alluded to as ”OWL 2”, was created by the [W3C OWL Working Group] (presently shut) and distributed in 2009, with a Second Edition distributed in 2012. OWL 2 is an expansion and update of the 2004 adaptation of OWL created by the [W3C Web Ontology Working Group] (presently shut) and distributed in 2004. The expectations that make up the OWL 2 detail incorporate a Document Overview, which fills in as a prologue to OWL 2, depicts the connection between OWL 1 and OWL 2.

OWL is an augmentation of RDF to defeat the inadequacies of the last option. To be sure, RDF experiences specific limits, which are, among others:

[1]http: www.ontoknowledge.org/Definedclas

- $rdfs : range$ characterizes the scope of qualities of a property paying little heed to the class concerned. For instance, it doesn't communicate that newborn children drink just milk while other age levels can drink pop.

- RDFS doesn't permit the articulation that two classes are disjoint. For instance, the classes of people are disjoint.

- RDFS doesn't permit you to make classes by a set blend of different classes (convergence, association, supplement). For instance, we need to build the Person class as the disjoint association of the classes of people.

- RDFS doesn't permit you to set a limitation on the quantity of events of qualities that a property can take. For instance, you can't say that an individual has precisely two guardians.

- RDFS doesn't permit characterizing specific qualities of the properties: transitivity (for instance isLarger-Than), uniqueness (for instance isTheFatherOf), converse property (for instance: eat is the reverse property of isMangaBy).

OWL ought to be based on RDFS while having XML punctuation. OWL permits classes to be characterized in a more mind boggling manner utilizing set (as well as intelligent) administrators like convergence, association, limitation, and so on), converse or transitive properties, or even limitations of cardinality on the properties, and this depends on a rationale of depiction. OWL likewise works with machine-level interoperability of web content more than whatever is as of now upheld by XML, RDF, and RDF Schema by furnishing extra jargon with formal semantics.

Also, OWL is blessed with three sub-dialects offering expanding limits of articulation and it is as indicated by the requirements, one picks the suitable language.

- OWL Lite is the simplest sub-language of OWL. It is intended for users who need a simple concept hierarchy. OWL Lite is suitable, for example, for rapid migrations from old thesauri.

- OWL DL is more complex than OWL Lite, allowing much more expressiveness. OWL DL is founded on descriptive logic (hence its name, OWL Description Logics), a field of research studying logic, and therefore giving OWL DL its adaptation to automated reasoning. Despite its relative complexity compared to OWL Lite, OWL-DL guarantees the completeness of reasoning (all inferences are calculable) and their decidability (their calculation is done in a finite period).

- OWL Full is the most perplexing variant of OWL, yet in addition the one that permits the most elevated level of expressiveness. OWL Full is expected for circumstances where have a significant degree of portrayal limit, regardless of whether it implies not having the option to ensure the fulfillment and the decidability of the computations connected to the metaphysics. Nonetheless, OWL Full offers intriguing systems, for example, the chance of broadening the default jargon of OWL.

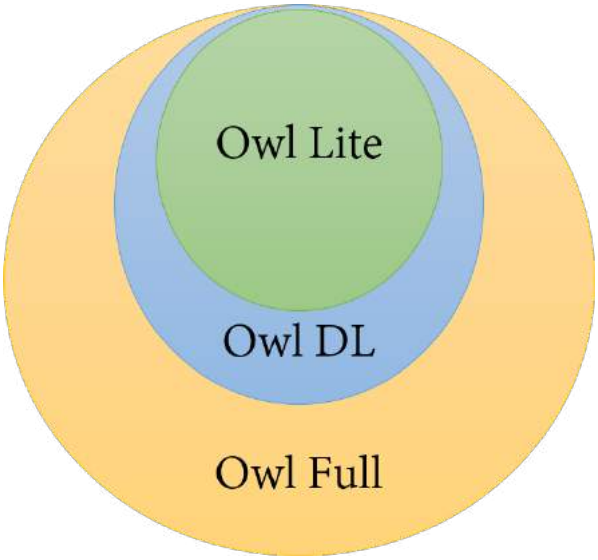


Figure 3.9: The layers of the OWL

There is a progressive reliance between these three sub-dialects: any legitimate OWL Lite metaphysics is additionally a substantial OWL DL cosmology, and any legitimate OWL DL philosophy is likewise a substantial OWL Full philosophy.

3.6.5 Description Logics(DL)

Depiction rationales (DL) is a group of formal information portrayal dialects. Numerous DLs are more expressive than propositional rationale however less expressive than first-request rationale. As opposed to the last option, the center thinking issues for DLs are (typically) decidable, and proficient choice strategies have been planned and carried out for these issues. There are general, spatial, worldly, spatiotemporal, and fluffy portrayal rationales, and every depiction rationale includes an alternate harmony between expressive power and thinking intricacy by supporting various arrangements of numerical constructors (305).

DLs are utilized in man-made consciousness to portray and reason about the pertinent ideas of an application space (known as expressed information). It is of specific significance in giving a consistent formalism to ontologies and the Semantic Web: OWL and its profiles

depend on DLs. The most striking utilization of DLs and OWL is in biomedical informatics where DL aids the codification of biomedical information.

3.7 Conclusion

The main objective of an ontology is to model a body of knowledge in a given domain. Thus, they play important role employment of formal representation of knowledge on a real-world subject and thus translate an explicit consensus. In the CBIR, ontologies may be used to, on the one hand, model Web resources from conceptual representations of the domains concerned and, on the other hand, with the objective of allowing automatic processing on them. The intelligent part is to mix two big fields together in such a way that, both fields affect each other and overcome problems and enhance their robustness and efficiency. In this chapter, we had detailed everything related to the ontology learning models as well as the ontology learning ”layer cake”.

4

Visual Relationship Extraction in Images and a Semantic Interpretation Ranking with Ontologies

4.1 Introduction

Today, with the rapid growth of a huge number of webly images, it becomes very challenging to capture, index, and analyze the relationships between objects even in only one image (306, 307, 308, 309, 310, 311). Given N objects and R predicates, a relation detection model has to examine ($O(N^2 \times R)$) relations. These would lead to a huge number of potential relation types in real-world applications. For example, there exist more than 75K relation types in the Visual Genome dataset (1).Because of that, three difficulties show up with time and they are considered as issues that should be addressed to assemble a solid model that will be utilized to extricate and semantically decipher the connection between objects in pictures (2, 3, 4), namely; long-tail problem (14), large intra-class divergence (16) and the semantic dependency or semantic gap (6, 7, 8).

The long-tail problem appears mostly in infrequent predicates between objects compared to the high occurrence of others in the same dataset. Intra-class divergence is where the same predicate (especially the spatial relationship predicates) can be used by a huge number of <object-subject> couple (such as on, under, etc.). Semantic dependency is solved only by extracting semantically the relationship between the triplet <object-predicate-subject>; High-level interpretation -namely a semantic relationship- between pairs of objects is needed. Taking an example in Figure 4.1, the VGD system detected a <woman-on-pants>, semantically, the prediction should be <woman-wearing-pants>.

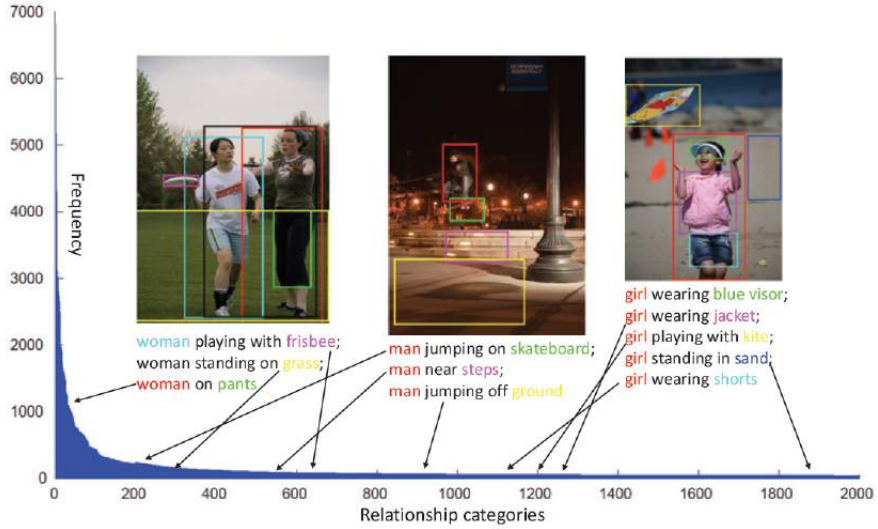


Figure 4.1: The long-tail label distribution of HCVRD dataset (4)

As illustrated in Figure 4.1, the authors in (4) only showed the top-2000 relationships because the tail is too long and can't be plotted at once. Three example images are shown, with webly-supervised model detected results. The shade of people and articles in the expressions relate to the shade of the jumping boxes. The arrows indicate the 'location' of the relationship in the label distribution. As we can see, most of the relationships lie in the tail. Some of them such as 'girl wearing blue visor' is not even in the top-2000. (4). To overcome the long tail problem, infrequent criteria of human-object relationships were targeted by HCVRD dataset. It has something like one human in the picture, and every human can have various associations with numerous items. It provides a rich set of predicates comprising about 927 categories.

In this chapter, we use the large-scale human-centric visual relationship detection dataset (HCVRD) proposed by (4) as an experimental example and it is not exclusively dedicated for it, but can be used with any other dataset that aims to be used the relationship between objects in images. The long tail is the common issue in all datasets, nonetheless, a dataset can't be good without being large and wide, and that causes falling into that problem. For that, the long-tail can be negligent in order to gain strong and robust results from using the benefits of large datasets. But, HCVRD is built in such a way that, the long-tail distribution issue and the zero-shot problem are maintained and solved already (see Figure 4.1). Due to that, we will address in this chapter solving the large intra-class divergence, and the semantic dependency or semantic gap.

The object extraction (or. relationship) and its semantic interpretation, was the target of many types of research till now (312, 313, 314, 315). Our search approach has been oriented

to solve large intra-class divergence and the semantic gaps with benefiting from the advances in the ontology. Ontologies are "an express particular of a conceptualization" (25). They ensure a mutual perspective of a specific space, just as a conventional model that is amiable to solo machine handling (26).

In any case, notwithstanding ontologies helped in understanding the semantic data from unstructured data(316, 317), for example, pictures or text (318), however the majority of the intriguing kinds of explores chipped away at hybridizing them with different calculations, for example, computerized reasoning calculations (319). In (320), the creators proposed a philosophy based picture comment driven by grouping utilizing HMAX highlights. The thought is (1) to prepare visual element classifiers and to construct a philosophy that can finely address the semantic data related with preparing pictures, and (2) to consolidate classifier yields and cosmology for picture explanation. To explain pictures, he characterized an enrollment worth of words in pictures. An assessment is ruined the enrollment esteem dependent on the certainty worth of classifiers and the semantic likeness between words. The enrollment esteem relies upon the word connections found in the cosmology that serve to choose an-documentation words. The acquired exploratory outcomes showed that the double-dealing of both classifier yields and metaphysics by assessing our proposed enrollment esteem empowers an improvement of picture comment.

The work in (321) introduced an original philosophy, dataset, and metaphysics driven profound learning approach for the arrangement of newsworthy occasion types in pictures. An enormous number of occasions related to an information base were utilized to recover a metaphysics that covers numerous conceivable true occasion types. The comparing huge scope dataset with 570,540 pictures permitted us to prepare incredible profound learning models. The outcomes on a few benchmarks have shown that the incorporation of organized data from a cosmology can further develop occasion grouping

The work in (322) presented another descriptor for pictures which permits the development of effective and conservative classifiers with great exactness on object classification acknowledgment. The prepared classes are chosen from a cosmology of visual ideas. The benefit of this descriptor is that it permits object-classification inquiries to be made against picture information bases utilizing proficient classifiers (productive at test time, for example, direct help vector machines, and permits these questions to be for novel classifications. In any event, when the portrayal is diminished to 200 bytes for each picture, grouping exactness on object classification acknowledgment is similar with the cutting edge (36% versus 42%), yet at significant degrees lower computational expense.

Figure 4.2 represents the strategies followed to construct an ontology from unstructured data, and the whole process is known as ontology learning layer cake (323). The acquisition

process extracts terms and their synonyms from input text needed to form concepts. After that, a process of relationship searching between those concepts is carried whether it is a taxonomic or non-taxonomic relationship. The last stage is to instantiate and extract general axioms.

Figure 4.2: Ontology learning layer cake (323)

Surveying the nature of cosmology procurement is a vital viewpoint. During the assessment, it is feasible to refine and redesign the whole cosmology learning process in the event of sudden resultant philosophy that doesn't fit with the particular prerequisites of clients (323). To avoid this issue, a pre-refined lexical database can be used such as the large lexical database "WordNet ontology" (324), or it can be used partially and refined to cover the user case of study. To summarize, the main contributions of this chapter include:

1. We use the HCVRD dataset that is built with an overcome of two important practical problems, the long-tail distribution issue, and the zero-shot problem.

2. We use the ontology learning layer cake strategy to build an ontological model that refines and models the acquisition of visual interpretation onto a lexical formal description.

3. We propose a statistical ranking module that aims to filter $false\ positives/negatives$ of class proposals. We start by making a comparison of object class proposals (i.e., probabilities of classification) that are obtained from the object detection module with the ontological formal description.

4. We propose two models that use the output statistical ranking module and work in parallel to further be combined to form the final output (i.e., classification of <human-predicate-object>); namely, semantic ontological module and visual relationship module. The semantic ontological module aims to achieve a high prediction' rank between semantically connecting objects in images. On the other hand, the visual relationship

module ranks the predicted relationship classes (i.e., predicates) by transferring the spatial relationship onto a high dimension spatial feature.

5. Finally, a demonstration of application scenarios is done in order to highlight the efficiency and the robustness of the proposed ideas.

4.2 Motivations and Proposals

An overview of our proposed semantic relationship detection (SRD) model is shown in Figure 4.3. Our model is divided into shared three sub-modules. The statistical ontology module, semantic relationship-HO ranking module visual relationship ranking module.

We apply the statistical ontology module during inference, we use it to detect all objects/relationships in the image. Then, we apply our new ranking modules to the detected class proposals, after that we eliminate the $false\ positive/negative$ class proposals. For the semantic relationship-HO ranking module, we build new features based on ontology semantic structure. Finally, for the visual relationship ranking module, we extract the bounding boxes to be further transformed onto highly geometric spatial coordinates. The results of the last two modules are used to classify the human-object relationships.

The feature extraction module is a stack of convolutional layers and max-pooling layers that have the same configuration as the VGG-16 (172) or the ResNet (153). The detection module is in the style of Faster-RCNN (325), which is used to detect the human and objects.

The bounding box that encompasses the detected human-object pair (i.e, contains both human and object), is sent to a deep metric learning module, which performs inference by finding the nearest-neighbor match in the web-crawled data amongst all triples sharing the same human and object labels; this determines the predicate category. The neighborhood distances are computed using the learned distance metric (i.e. in the feature space). The three sub-modules can be learned in an end-to-end manner. For efficiency, the feature map generated by the feature extraction module is shared as input to the followed two modules. We use the VGG-16 (172) network as a basic building block for our model. We discuss the detection module and the distance metric learning module in more detail in the following sections.

4.3 Problem Formulation

4.3.1 Ontological model

For the artificial intelligence team group, a philosophy is an express particular of conceptualizations (25). It is a depiction of a bunch of illustrative natives with which to demonstrate

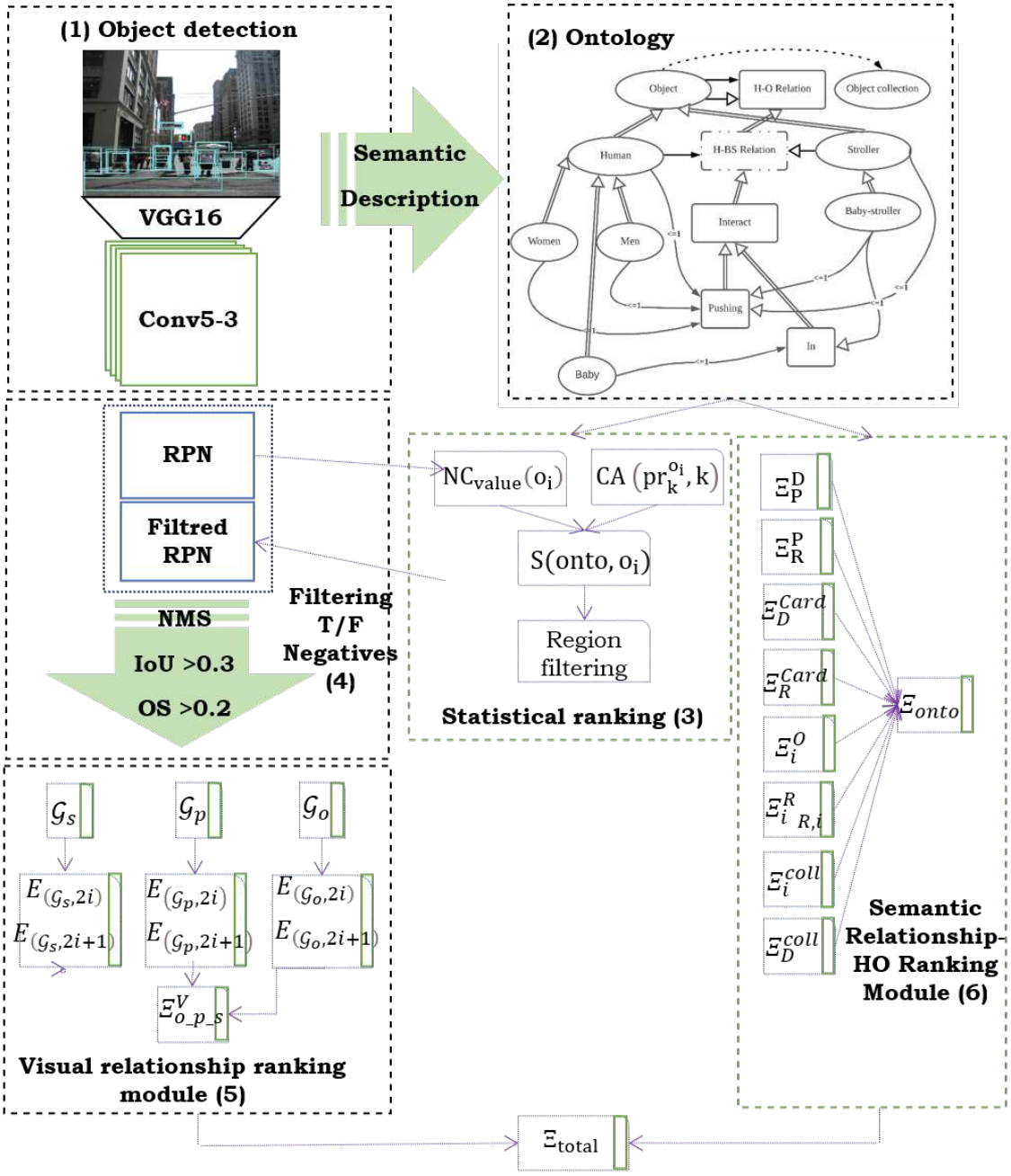


Figure 4.3: Content visual relationship and semantic interpretation ranking with ontologies applied to an example image. (1) represents the object detection module for an input image. (2) is the ontology building module that contains background knowledge of the detected objects and their relations. (3) uses the outputs of the object detection module and ontology module. It represents the statistical ranking module that aims at filtering false negatives/positives in (4). (5) is the output of (4) where a visual relationship ranking module is done based on transforming the spatial features onto a high dimension. (6) the output of the ontology background knowledge is used to rank the semantic relationship between $< human - object >$ pairs

a theoretical model of an information space.

To minimize the gap between visual and semantic perception, we provide the ontological model as part of our system. This model provides useful background knowledge about image objects and their relations (an example is shown in Figure 4.3(2)) where it is constructed based on the ontology description presented in this section (a joint credit ontology description example is illustrated in Figure 4.4).

Taking some detected objects in an image, the ontology model will provide *Object class*, $OO - Relations$, and *Object Collection*. Every one of these Object classes is a quick subclass of Object. The philosophy model likewise will give semantic information ($\mathcal{H}_R$ and $\mathcal{S}_R$) that contains foundation information about the semantic pecking orders of item classes and connection classes, and the requirements on connection classes between them.Formally, we characterize a cosmology $O = \langle C_o, R, I, A \rangle$ as follows (326):

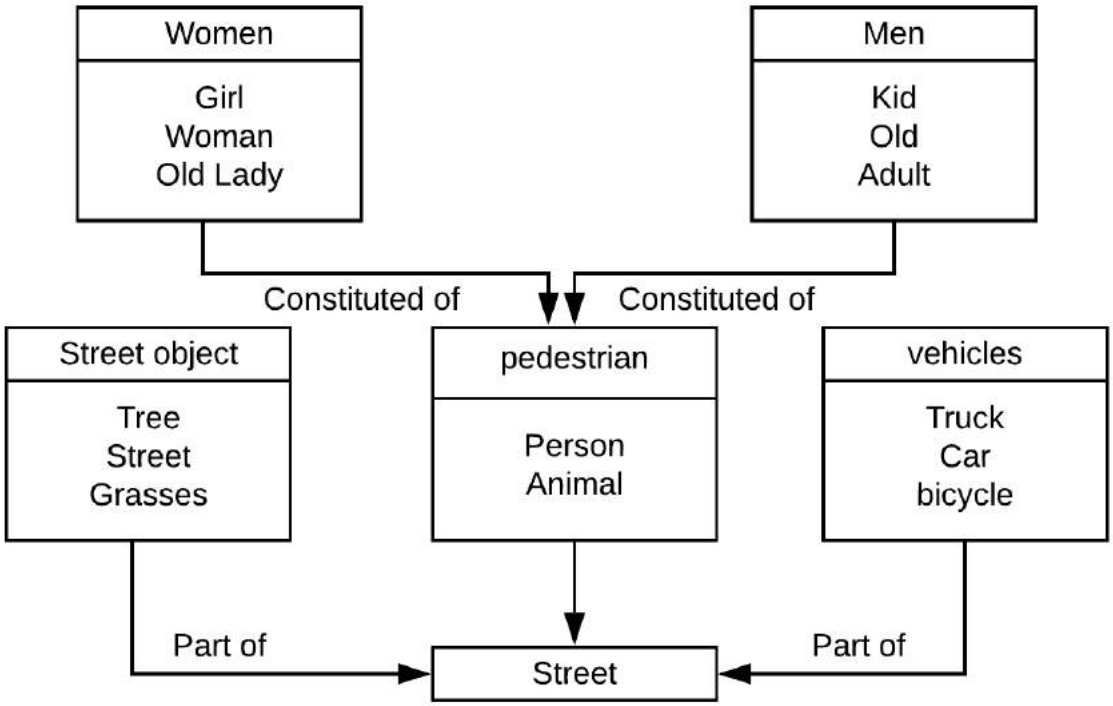


Figure 4.4: The joint credit ontology; illustrative example

- $\mathcal{C}_o = \{c_{o1}, ...c_{on}\}$ is the arrangement of n ideas (classes) with the end goal that each c_{oi} has a bunch of k properties (ascribes) $\mathcal{O}_i = \{o_1, ...o_k\}$. Such as: $\mathcal{C}_o$ ={$Womem$, Man, $Pedestrian$, $Street\ Object$, $Vehicles$}.
- $\mathcal{R}$ is the arrangement of double relations between components of $\mathcal{C}_o$ which comprises of two subsets

  - $\mathcal{H}_R$ which depicts the legacy relations among ideas and here and there can be addressed as an adage. For example, "Is-a" relations address $\mathcal{H}_R$ and are comparable to subsumption sayings, e.g., Man $\subseteq$ Person.
  - $\mathcal{S}_R$ which depicts semantic relations between ideas. That is, every connection $c_{oi}S_Rc_{oj} \in \mathcal{S}_R$ has c_{oi} as a space and c_{oj} as a reach. For example, $\mathcal{S}_R$ = Part of (Street−object, Street).

- $\mathcal{I}$ is the arrangement of occasions, addressing the information base. For example, the occurrence "$\mathcal{P}_{vehicles}$" = $\{truck; Car; Bicycle\}$.

- $\mathcal{A}$ is the arrangement of the maxims of the cosmology. $\mathcal{A}$ comprises of limitations on the space of the cosmology that include $\mathcal{C}_o$, $\mathcal{R}$ and $\mathcal{I}$.

Sayings are with type of $A \equiv B$ (A and B are same), $R1 \subseteq R2$ ($R1$ is a sub-property of $R2$), $R1(x; y)$ (x is identified with y by the connection $R1$), $A(x)$ (x is of type A), and so on Where A and B are ideas, $R1$ and $R2$ are relations and x and y are occasions.

Note that, models with profound rich various leveled data and high limitations are consistently solid as they convey more semantic data. We present the accompanying foundation information in our ontological model:

- **Subsumption** is a relationship where one class is a subclass of another, indicated as $A \sqsubseteq B$. E.g., Old-woman is a subclass of females.

- **Domain/range** requirements state the area or reach object class of a connection class, signified as $domain(C)$ and $range(C)$. E.g., in Figure 4.3, the space and the scope of pushing-of should be men and $baby - stroller$ separately.

- **Cardinality constraints** limit the most extreme number of relations of a specific connection class that an article can have, where the item's class is a space/scope of the connection class. E.g., in Figure 4.3(1), $Men \xrightarrow{\leq 1} pushing - of$ implies that $woman$ can have all things considered one $Pushing - of$ connection.

- **Collection** alludes to a bunch of picture objects having a place with a similar item class, mean as $collection(C)$.

4.3.2 Mathematically Problem Formulation

Presently, given a bunch of identified items $\{O_i\}$ in the info picture, we make one article hub o_i for each item O_i, and one connection hub $r_{i,j}$ for each article pair $< O_i, O_j >$ that has a comparing connection class of the aide metaphysics (e.g., object pair $< men1$, $baby -$

$stroller1 >$ relating to class $H - BS\ Relation$). A marking is thinking about as plausible when it fulfills:

1. For a set of class assignments $\mathcal{C}(o_i)$ for object o_i, there is $\mathcal{C}(o_i) = \{C_o \mid C_o \sqsubseteq C_g(O_i)\}$, where $C_g(O_i)$ is the generic class of object O_i (e.g., Men for object $men1$).

2. For a set of class assignments $\mathcal{C}(r_{i,j})$ for relation node $r_{i,j}$, there is $\mathcal{C}(r_{i,j}) = \{C_r \mid C_r \sqsubseteq C_g(O_i, O_j)\}$, where $C_g(O_i, O_j)$ is the corresponding relation class (e.g., $H - BS\ Relation$).

The best possible marking is needed to (1) fulfill the cosmology adages and limitations, (2) be profoundly and luxuriously enlightening, and (3) boost the position of the class task of every trio $< Human - object >$ concerning their visual appearance. We anticipate the ideal marking by limiting the position Ξ_{total} concerning an info picture and a metaphysics model:

$$\Xi_{total} = \Xi_{onto} + \Xi_{ops}^{V} \tag{4.1}$$

$Equation1$ Represents the sum of the Semantic Relationship-HO Ranking and the Visual relationship ranking respectively. In the following sessions, we will explain and detail each part.

4.4 Statistical Ontology Module

The article (or., subject) recognition module structure is indistinguishable from that of the Faster-RCNN (325). Taking the yield of the element extraction module ($Conv5$ 3 include map) as information, the Region Proposal Network (RPN) (325) is utilized to produce object recommendations. During preparing, we extricate highlights with $RoIPool$ (325, 327) for each article proposition, trailed by the jumping box relapse misfortune and cross-entropy misfortune to become familiar with the indicator/classifier. The learning rate is introduced to 0.0001 and diminished by an element of 10 after each 5 age. During induction, we utilize this module to identify all articles (or., subjects) in the pictures. Then, at that point, we apply the proposed factual metaphysics module to kill the $false\ positive/negative$ instances of items (or., subjects) anticipated proposition. As result, we acquire sifted class recommendations that are utilized for the remainder of the cycle. From that point forward, we apply non-greatest concealment (NMS) on the separated class recommendations with the IoU (Intersection of Union) edge 0.3 and objectiveness scores higher than 0.2.

In order to achieve the filtered class proposals, we apply two techniques that are used in an end-to-end manner, namely, C/NC ranking and Contrastive analysis ranking. In literature, both techniques were used separately to achieve the relative objectives for the tasks of Neural

Language Processing (NLP) domain-based ontology. In this chapter, we take the first use of those techniques in an end-to-end manner on content-based image retrieval ($CBIR$). We adopt the same concepts and we apply to them a normalization of coefficients in order to make them compatible with -class probability- (P_c) obtained from $RoIPool$. In the following subsections, we describe the statistical ontology module techniques, i.e., C/NC ranking, and Contrastive analysis ranking:

4.4.1 C/NC ranking

C/NC value is used to capture the whole context relevant to describing an input image to limit the number of object class proposals. Those proposals are used to calculate the C/NC value that is a combination of C_{Value} and NC_{Value}. C_{Value} tends to find a group of the class proposals that are with high prediction probabilities. Whereas, NC_{Value} is a modification in C_{Value} that considers the high predicted class proposals and tries to find the most frequently appeared of those class proposals. Formally, C_{Value} can be calculated as:

$$C_{value}(o_i) = \begin{cases} log_2|o_i| \times f(o_i)\ only\ if \sum_{i=1}^{card(C)} o_i \models C_i \\ log_2|o_i| \times f(o_i) - \frac{1}{C(o_i)} \sum_{k=1}^{Card(o_i)} f(o_k)\ only\ if\ \sum_{i=1}^{card(C)} o_i \models \neg C_i \end{cases} \tag{4.2}$$

In equation 2, $f(o_i)$ counts the frequency of o_i in the set of objects. a$C(o_i)$ count the number of the class proposals of o_i with high probabilities. When C esteem rank is found, the subsequent stage is to join logical data about the entire picture. For that, a normalization of the object frequency (that belongs to the same image) is done based on the following equation.

$$weight(o_i) = \frac{f(o_i)}{Card(\mathcal{C})} \tag{4.3}$$

A weight esteem is then added into C_{value} to get the NC_{value}. Numerically, it tends to be composed by the accompanying equation, where α is an enhancement factor.

$$\mathcal{NC}_{value}(o_i) = \alpha \times \mathcal{C}_{value}(o_i) + (1 - \alpha) \times weight(o_i) \tag{4.4}$$

Taking an example of an input image with five objects that are found with high probabilities to be labeled as "car" and similarly two objects that are found with high probabilities to be labeled as "table". The idea here is to eliminate this second prediction (i.e., the probabilities that an object is highly labeled as "table") since the whole context of the input image is not homogeneous, i.e., the probability that a "table" can be found in the same place with a "car" is mostly weak.

4.4.2 Contrastive analysis ranking

The contrastive analysis is used to eliminate terms that are not relevant to the context. Here, we adopt the same concept and transform it to serve our goals. In order to filter the irrelevant class proposals of each object, two measures are calculated, namely; proposal relevance (relevant) and proposal consensus (non-relevant). Separating guarantees that the recommendations that are more applicable to the unique situation, will remain. Area importance is utilized to quantify the explicitness of a term concerning its reach. Formally, Given a set of class-proposal of every object $o_i \models C = \{pr_0^i, pr_1^i, \dots\ pr_N^i\}$, and a list of contrastive class proposal $\{S_0^{pr}, S_1^{pr}, S_2^{pr}, \dots S_m^{pr}\}$ is calculated using the normalized cosine similarity illustrated on *Equation*5 and *Equation*6. the m value with small distances is sorted to represent the contrastive class proposal.

$$\cos(\theta) = \frac{o_i^\top \times o_j}{\|o_i\| \times \|o_j\|} \tag{4.5}$$

$$Similarity = \frac{\sum_{j=0}^{N} o_i(pr_j^i) \times o_k \times (pr_j^k)}{\sqrt{\sum_{j=0}^{N} o_i(pr_j^i)^2} \times \sqrt{\sum_{j=0}^{N} o_k(pr_j^k)^2}} \tag{4.6}$$

For a class proposal pr_k^0 (0 is the object in interest). The proposal relevance PR is measured as:

$$PR\left(pr_k^0, k\right) = \frac{P\left(pr_k^o | S_k^{pr}\right)}{\sum_{i=1v}^{m} P\left(pr_k^o | S_i^{pr}\right)} \tag{4.7}$$

Where $P(pr_k^o|S_k^{pr})$ and $P(pr_k^o|S_i^{pr})$ are the probabilities of finding a class proposal pr_k^i in the contrastive domain S_k^{pr} and contrastive class proposal S_i^{pr}, respectively.

Then again, proposition agreement is utilized to find the class recommendations that show up in a few class recommendations of the contrastive area S_k^{pr}. It very well may be determined as:

$$PC(pr_k^o) = \sum_{k=1}^{m} P\left(pr_k^o | pr_i^k\right) \times \log\left(\frac{1}{P\left(pr_k^o | S_k^{pr}\right)}\right) \tag{4.8}$$

where $P\left(pr_k^o|S_k^{pr}\right)$ stands for the probability of a class proposal pr_k^0 in the contrastive domain S_k^{pr}. The two measures are then integrated together using linear combination formula that is stated as:

$$CA\left(pr_k^o, k\right) = \alpha \times PR\left(pr_k^o, k\right) + (1-\alpha) \times PC\left(pr_k^o, k\right) \tag{4.9}$$

α is an experimental parameter between 0 and 1. After calculating the scores for each class proposal in the contrastive domain. A selection is done based on high scores obtained.

$$S(onto, o_i) = CA\left(pr_k^{o_i}, k\right) + NC_{value}(o_i) \tag{4.10}$$

After applying statistical techniques; filtered boxes are obtained to be grouped as pairs, i.e., $< human, object >$, and a bounding box that fully contains the human and object boxes are associated with each pair. These "union" bounding boxes are (separately and individually) the input of the Semantic Ontology module and the Visual relationship detection module, where each module is dedicated to calculating the ranks that are used as features in visual relationship final prediction.

4.5 Semantic Relationship-HO Ranking Module

We define the ranks based on background knowledge in the guide ontology.

4.5.1 Domain/range ranking

We will rank an assignment of a relation between an object (domain) and a subject (range) if it exists an ontological edge between them. In contrast, a strong penalty (infinity) is assigned to the rank score in the case of non-linked objects.

$$(o_i; r_{i,j}) \rightsquigarrow \Xi_P^D = \omega_{D,P} \; only \; if \; o_i \models C_o \; and \; r_{i,j} \models C_r$$

$$(o_i; r_{i,j}) \rightsquigarrow \Xi_P^D = \infty \; only \; if \; o_i \models \neg C_o \; or \; r_{i,j} \models \neg C_r$$

$$(r_{i,j}; o_j) \rightsquigarrow \Xi_R^P = \omega_{P,R} \; only \; if \; r_{i,j} \models C_r \; and \; o_j \models C_o$$

$$(r_{i,j}; o_j) \rightsquigarrow \Xi_R^P = \infty \; only \; if \; r_{i,j} \models \neg C_r \; or \; o_j \models \neg C_o$$

4.5.2 Cardinality ranking

The cardinality of domain-range predicate plays an important role at the ontology level. Taking pair of $< Human - Object >$, it is important to note that, there is a restriction in number of instances. For example, when taking two objects *baby stroller*1 and *baby stroller*2; and a subject *men*1. We can't assign a predicate *pushing* to the pair $<$ *baby stroller*1, $men - 1 >$ and in the same time to $<$ *baby stroller*2, *men*1 $>$. If one *predicate* ranked high and classified for a $< Human - object >$ pair, the ranking of the same *predicate*

for others pairs (or, if two *predicates* are assigned to the same *object* $-orhuman-$) will give a high penalty ($men1$ can push only one *baby stroller*1) or subject (a *baby stroller*1 is pushed by only one $men1$).

$$(r_{i,j}; r_{i,k}) \rightsquigarrow \Xi_D^{Card} = \omega_{card,D} \; only \; if \; r_{i,j} \models C_1 \; and \; r_{i,k} \models C_2 \; and \; C_1 \models C_2 \models C_r$$

$$(r_{i,j}; r_{i,k}) \rightsquigarrow \Xi_D^{Card} = \infty \; only \; if \; (r_{i,j} \models C_1 \; and \; r_{i,k} \models C_2) \; and \; C_1 \models \neg C_2$$

$$(r_{i,j}; r_{k,j}) \rightsquigarrow \Xi_R^{Card} = \omega_{card,R} \; only \; if \; r_{i,j} \models C_1 \; and \; r_{k,j} \models C_2 \; and \; C_1 \models C_2 \models C_r$$

$$(r_{i,j}; r_{k,j}) \rightsquigarrow \Xi_R^{Card} = \infty \; only \; if \; (r_{i,j} \models C_1 \; and \; r_{k,j} \models C_2) \; and \; C_1 \models \neg C_2$$

4.5.3 Depth information ranking

As we mentioned before, a feasible and strong ontology is built with more specific and more deep rich information. Depth information refers to the depth of the assigned class in the subclass tree. A high rank is given to assignments with high depth since it is more informative and thus may be more attractive to the users. In contrast, assignments with small depth will be penalized to be avoided since only a little information revealed. Therefore, we give a high rank for each object o_i or predicate $r_{i,j}$ with high depth information:

$$(o_i) \rightsquigarrow \Xi_i^O = \omega_{O,i} \times deep(o_i) \; only \; if \; o_i \models C_o$$

$$(r_{i,j}) \rightsquigarrow \Xi_i^R = \omega_{R,i} \times deep(r_{i,j}) \; only \; if \; r_{i,j} \models C_r$$

4.5.4 Collection ranking

Collection ranking defines the type of objects that belong to the same class assignment. Intuitively, it reflects the long tail problem since collections with small sizes are not preferable. And as we are using HCVRD dataset (4); a long-tail problem was solved and that is leading to collections with the larger size.

Our ontology model is built with deeper information to be more informative. A high rank is given to a deep and large collection. We create a virtual and instant edge between object nodes when they belong to the same object class. An example of collection ranking is illustrated in Figure 4.5. If two persons (i.e., two objects) are relatively close to each other's based on the spatial visual appearance (see next session) and based on the strong visual observation, they may be labeled as $Lay-down$ on $Grass$, it is very normal to mark the third person with $Sleeper$ rather than $Worker$ and the whole image a well may be classed as $meadow$ rather than $building_area$.

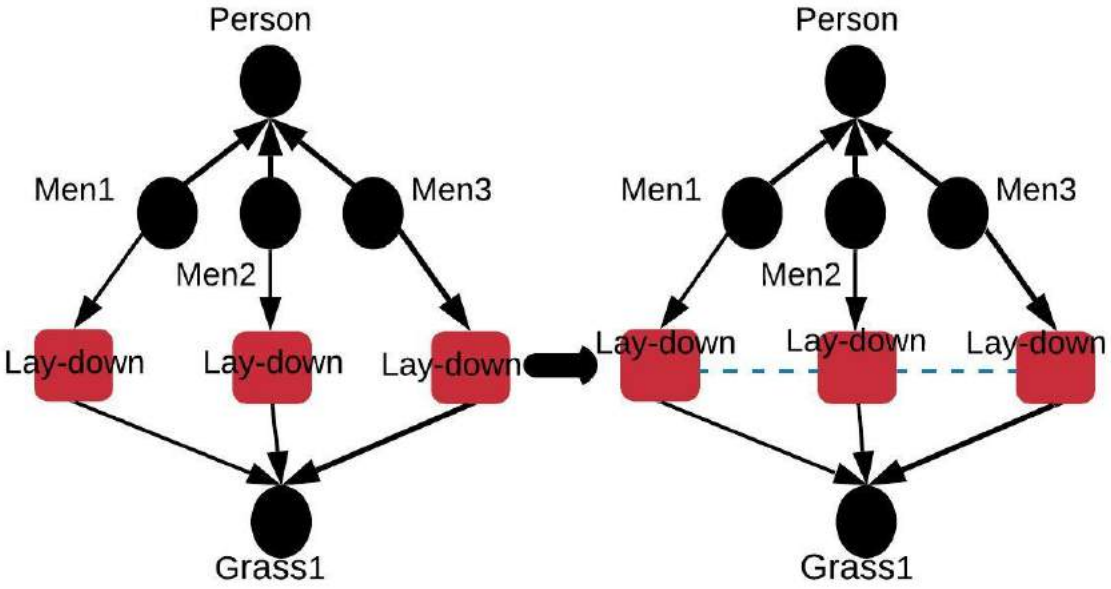


Figure 4.5: Collection ranking strategy

$$(o_i, o_j) \rightsquigarrow \Xi_D^{coll} = \omega_{coll} \times \tfrac{2}{N-1} \times deep(coll(C_o)) \; only \; if \; o_i \models C_1 \; and \; o_j \models C_2 \; and \; C_1 \models C_2 \models C_r$$

$$(o_i, o_j) \rightsquigarrow \Xi_D^{coll} = \infty \; only \; if \; (o_i \models C_1 \; and \; o_j \models C_2) \; and \; C_1 \models \neg C_2$$

where ω_{coll} is a weight, and $\frac{2}{N-1}$ is a standardization factor, N addressing the quantity of article hubs that can be conceivably named with C_o. At long last, the positioning based philosophy requirement is the amount of:

$$\Xi_{onto} = \Xi_P^D + \Xi_R^P + \Xi_D^{Card} + \Xi_R^{Card} + \Xi_i^O + \Xi_i^R + \Xi_D^{coll} \tag{4.11}$$

4.6 Visual relationship ranking module

4.6.1 Visual Feature Extraction

For various branches in our organization, we require their comparing highlights. For subject and article branches, we feed the common convolutional include map and the created object recommendations to a $ROI - pooling$ layer. We get the subject and article visual element vectors of size $512 - 7 - 7$ and straighten them. Through two $512 - dimensional \; FC$ layers, we acquire $512 - dimensional$ subject and item visual element vectors.

4.6.2 High dimension Spatial features

Predicate prediction performances can be enforced by combining spatial location with visual appearances. Works proved its efficiency by using object and subject individual locations (328) or by their union (predicate location) (325, 329). In this chapter, we will use high dimension geometric spatial location and transform them into features (330). We use the same vector dimensional as (330).

Let $\mathfrak{D} \in \mathcal{N}^*$ which is the individual feature dimension for each component, <Human, Predicate, Object> $< H, P, O >$, be a feature triplet that represents the final decision visual relationship of object-predicate-subject. o_s, o_o, o_{pr} which is the triplet bounding boxes that are the output the object detection phase. As result, $o_{hu} = [x_{hu}, y_{hu}, w_{hu}, h_{hu}]$, $o_o = [x_o, y_o, w_o, h_o]$, and $o_{pr} = [x_{pr}, y_{pr}, w_{pr}, h_{pr}]$. We take the vectors $\mathcal{G}_{hu}, \mathcal{G}_o, \mathcal{G}_{pr}$ with dimensions of 6, 8, and 6 to represent geometric spatial location. Finally, the vectors are then used in the high dimension transformation.

As mentioned before, the aim is not to capture the triplet bounding boxing only as a feature. But, it is to transform them into a high dimension features. *Equation*12, *Equation*13, and *Equation*14 represent the geometric spatial location for the subject, object, and predicate respectively. We transform them into high dimensions using *Equation*15 and *Equation*16 respectively.

$$\mathcal{G}_{hu} = \left(\frac{x_o - x_{hu}}{w_{hu}}, \frac{y_o - y_{hu}}{h_{hu}}, \log \frac{w_o}{w_{hu}}, \log \frac{h_o}{h_{hu}}, x_{o,central}, y_{o,central} \right) \tag{4.12}$$

$$\mathcal{G}_o = \left(\frac{x_{hu} - x_o}{w_o}, \frac{y_{hu} - y_o}{h_o}, \log \frac{w_{hu}}{w_o}, \log \frac{h_{hu}}{h_o}, x_{hu,central}, y_{hu,central} \right) \tag{4.13}$$

$$\mathcal{G}_{pr} = \left(\frac{x_{hu} - x_{pr}}{w_{pr}}, \frac{y_{hu} - y_{pr}}{h_{pr}}, \log \frac{w_{hu}}{w_{pr}}, \log \frac{h_{hu}}{h_{pr}} \frac{x_o - x_{pr}}{w_{pr}} \frac{y_o - y_{pr}}{h_{pr}}, \log \frac{w_o}{w_{pr}}, \log \frac{h_o}{h_{pr}} \right) \tag{4.14}$$

Note that the first two components $x_{hu} - x_o$ and $y_{hu} - y_o$ represent the translation between subject and object that are normalized by the bounding box' weight w_o and height h_o respectively. the ratios between the contradictory weights and heights of the two boxes are also calculated which are $\log \frac{w_{\cdot}}{w_{\cdot}}$ and $\log \frac{h_{\cdot}}{h_{\cdot}}$. $x_{..,central}, y_{..,central}$ are the central point coordinate of the detected bounding box. (the same process is applied for the object and the predicate equations).

After obtaining $\mathcal{G}_s$, $\mathcal{G}_o$ and $\mathcal{G}_{pr}$, we transform them into high dimensions using *Equation*15 And *Equation*16 respectively, where D is the dimension of the model (we use $D = 32$). We repeat the same process and each time, we replace $\mathcal{G}$ with $\mathcal{G}_s$, $\mathcal{G}_o$ and $\mathcal{G}_{pr}$ respectively and separately.

$$\mathcal{E}_{\mathcal{G},2i} = \sin\left(\frac{\mathcal{G}}{10000^{\frac{2i}{D}}}\right), i = 0, ..., \left(\frac{D}{2} - 1\right), i \in \mathbb{N} \tag{4.15}$$

$$\mathcal{E}_{\mathcal{G},2i+1} = \cos\left(\frac{\mathcal{G}}{10000^{\frac{2i+1}{D}}}\right), i = 0, ..., \left(\frac{D}{2} - 1\right), i \in \mathbb{N} \tag{4.16}$$

After obtaining the geometric transformation of each bounding box spatial characteristic; a probability of each object is used to rank the visual relationship proposal and it is given by the ontological annotation that is presented as follows:

$$(o_i) \rightsquigarrow \Xi^V_{ops} = \omega_{ops} \times Prob(o_i \models C_1) + \omega_{sub} \times Prob(o_j \models C_2) + \omega_{pr} \times Prob(r_{i,j} \models C_r) \; only \; if \; o_i \models C_1 \; and \; o_j \models C_2 \; and \; r_{i,j} \models C_r$$

4.7 Tools and Experimental Results

To validate the effectiveness of the proposed scheme, we must answer this question: Q-Is the rich background knowledge offered by the ontology effective for the visual relationship detection? To answer this question, we evaluate the proposed methods on a newly released dataset HCVRD, which is collected from the Large Visual Genome dataset and other web images. Based on the extracted objects and relationships, we build a connected ontology using the WordNet prototype.

4.7.1 Datasets, Metrics, and Evaluation Setup

In order to evaluate our contributions, some benchmarks and tools are used. This sub-section is dedicated to illustrating their descriptions as well as their characteristics.

1. **HCVRD** (4): is a new dataset that considers the relationship between humans and objects in images. The major advantage of using HDRD is that it is built with some constraints that aim to avoid falling into the long-tail distribution issue and the zero-shot problem. HCVRD was gathered from the huge Visual Genome dataset (for example search of pictures that contain people). Likewise, an advantageous part of 788,160 pictures drawn from the best 100 picture query items for every relationship triple. Altogether, there are 52,855 pictures with 1,824 article classifications and 927 predicates. 256,550 connections examples with 9,852 non zero-shot relationship types and 18,471 zero-shot connection ships types. There are on normal 10.63 predicates per object class. 31,586 pictures are utilized for preparing and two test parts are utilized. The principal test split contains 10,000 pictures where every one of the connections happen in the preparation set. Another test split incorporates every one of the zero-shot connections.

2. **VRD:** The Visual Relationship Detection with Language Priors (VRD)(331) is a dataset that recently accomplished tint progress in visual connections recognition. It contains 5000 pictures with 37,993 thousand relationships,100 object classes, and 70 predicate classifications interfacing those items together. Predicates are ordered into the 5 after types: Action, Spatial, Preposition, Comparative, Verb.

3. **VG**: Visual Genome (VG) (332) is enormous dataset. It contains 99,658 pictures with 200 article classes and 100 predicates. There are absolutely 1,174,692 connection cases among 19,237 one of a kind trios. The default split contains 73,801 pictures for preparing and 25,857 pictures for testing.

4. **VGG-16 and Faster-RCNN**: The feature extraction module is a stack of convolutional layers and $max - pooling$ layers which have the same configuration as the $VGG-16$ (172). The detection module is in the style of $Faster-RCNN$ (325), which is utilized to distinguish the item and human subject (in its sub-classification).

5. **WordNet**: WordNet ontology (324, 333)is a lexical data set of semantic relations between words in excess of 200 dialects. The power of wordnet is beyond its ability to find the meaning of sentences using semantic relations (i.e. synonyms, hyponyms, hypernyms, meronyms, holonyms, entailments, etc.). With the help of WordNet, a computer program will be able to identify relationship and object classes. A demonstration is given in the next sections.

4.7.2 Semantic Relationship-HO Ranking Module Evaluation

WordNet has been utilized for various purposes in data frameworks, including word-sense disambiguation, data recovery, programmed relationship grouping. For us, WordNet ontology is used as a background to support the visual information extracted from the deep learning process. It provides us a strong knowledge about the relationships between detected objects for automatic image understanding. The following figure 4.6 is OntoGraph of WordNet OWL file extracted from "Protegé" that prototype the last version of the ontology. Due to that, the use of any sub-set of WordNet ontology will be structured with the head classes named individual, list, property, and resource. Figure 4.7 depicts the WordNet main classes, list, property, and resource. This figure is extracted from the "Protegé" OntoGraph library. For more detail and as an example, $has-intersection-of$ and $has-subclass$ represent the general form of two axioms. Figure 4.8 represents an example of English WordNet Interpretation for a triplet $Person-to-Person$ in the cases; adverbs and adjectives. WordNet considers for each entry the list of semantic relationship interpretations in term of synonyms, hyponyms,

hypernyms, meronyms, holonyms, entailments, etc. (if exist). It also offers a list of similar interpretations to that entry.

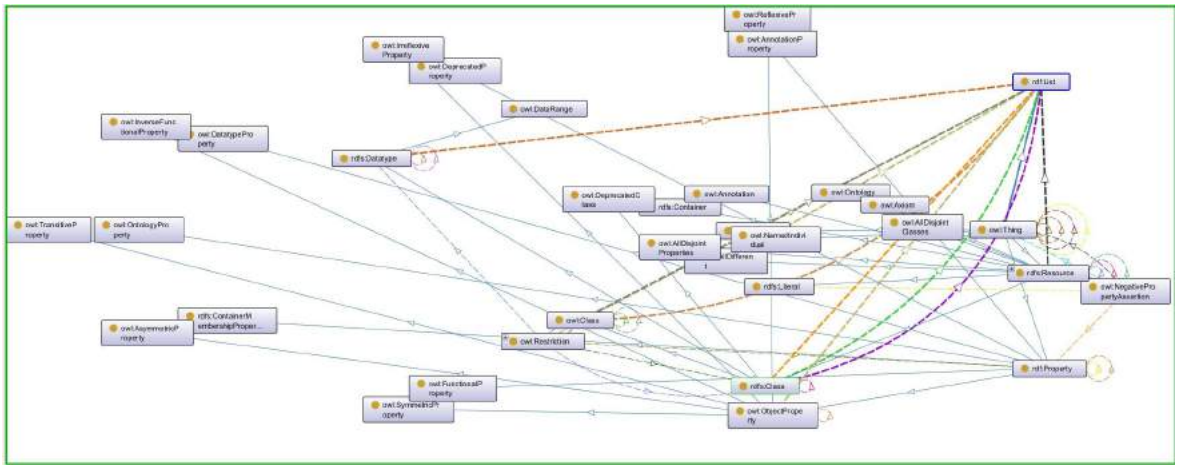


Figure 4.6: WordNet main classes, list, property and resource. This figure is extracted from "Protegé" OntoGraph library

To assess our work, we utilize 2000 distinct pictures from each dataset, HCVRD, VG, and VRD. 1200 unique pictures from each dataset are utilized as preparing information, and 800 distinct pictures from each dataset are utilized as test information.

All pictures were rescaled and the picture preprocessing methods were applied. This incorporates arbitrary editing of the picture just as irregular splendor, immersion, shade, and differentiation bending to stay away from overfitting. We pick the classifications: individual, road, vehicles, telephone, buggy. The relationship location between classes of each dataset is utilized to build the ontologies. In our model, we think about the separated predicates and connections between objects as information and improve it with a manual explanation for what we called profound rich data of between intra objects. The manual explanation is inside the metaphysics development, we utilized the depictions of ideas and expressions of the enormous lexical dataset "WordNet".

For the measurable ontological module, the article discovery module structure is indistinguishable from that of the Faster-RCNN. Taking the yield of the element extraction module (Conv5 3 component map) as info, the Region Proposal Network (RPN) is utilized to create object recommendations. During preparing, we separate elements with RoIPool for each item proposition, trailed by the bouncing box relapse misfortune and cross-entropy misfortune to gain proficiency with the finder/classifier. The learning rate is introduced to 0,0001 and diminished by an element of 10 after each 5 age. During deduction, we utilize this module to identify all items (or., subjects) in the pictures. Then, at that point, we apply the proposed factual cosmology module to dispose of the bogus positive=negative instances of anticipated

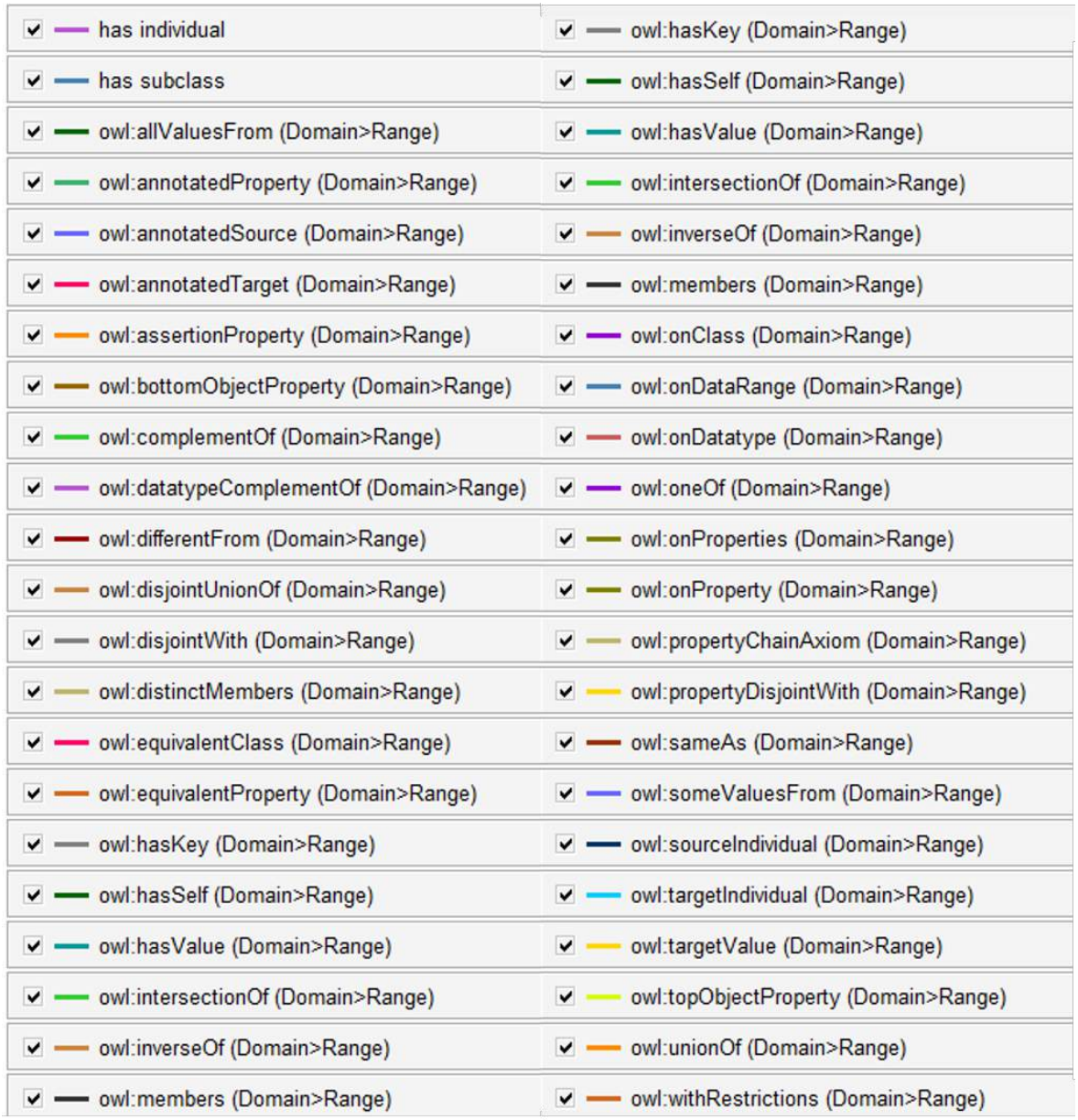


Figure 4.7: WordNet main classes, list, property and resource. This figure is extracted from "Protegé" OntoGraph library

recommendations. As result, we acquire separated class proposition that are utilized for the remainder of the cycle. From that point forward, we apply non-most extreme concealment (NMS) on the sifted class recommendations with the IoU (Intersection of Union) limit 0,3 and objectiveness scores higher than 0,2.

Or on the other hand the Visual Feature Extraction, we use VGG-16 to remove for various branches in our organization, their relating highlights. For subject and article branches, we feed the common convolutional include map and the produced object proposition to a ROI-

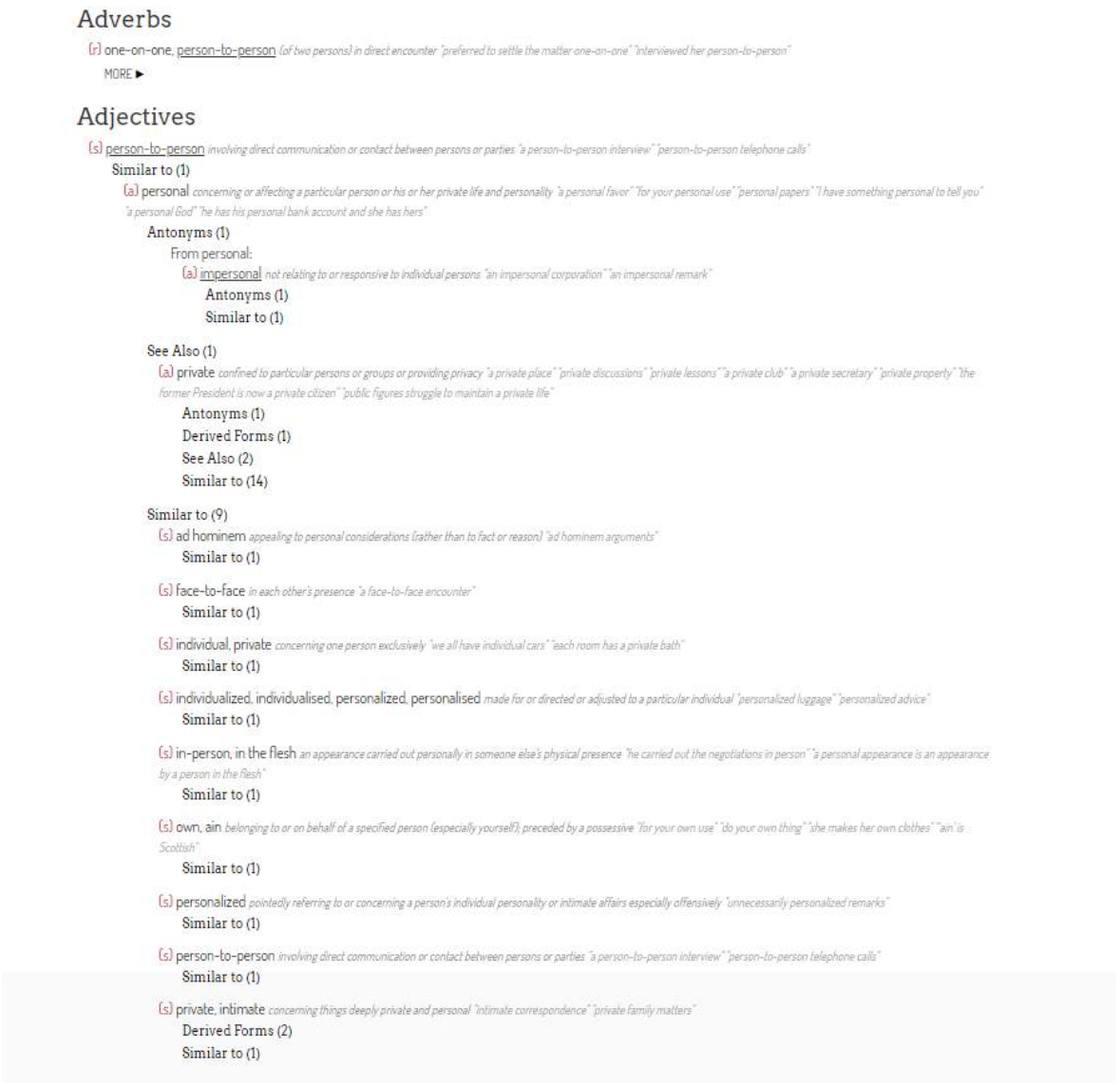


Figure 4.8: An example of English WordNet Interpretation for a triplet Person-to-Person in the cases; adverbs and adjectives

pooling layer. We get the subject and article visual element vectors of size 512_7_7 and level them. Through two 512_dimensional FC layers, we acquire 512_dimensional subject and item visual component vectors.

4.7.2.1 Manual Annotation Evaluation

In the following, we will show how manual annotation work for an input image that is captured by VG dataset. Normally, we use the relationships detected by the systems and extend them manually, but this image has no relationship detection by VG system. For that, and for similar situations, we develop our own annotation starting from zero. After

obtaining list of annotations, we use the lexical dataset WordNet to build for each object, predicate, and subject. The latter process is done using the ontology learning layer cake, we start by extracting terms and their synonyms to form concepts. After that, a process of relationship searching between those concepts is carried whether it is a taxonomic or non-taxonomic relationship. The last stage is to instantiate and extract for each object, subject, and predicate, their corresponding list of synonyms, descriptions and characteristics. An example can be seen on annexes session **??**, we showed the XML code for object term "wheel".

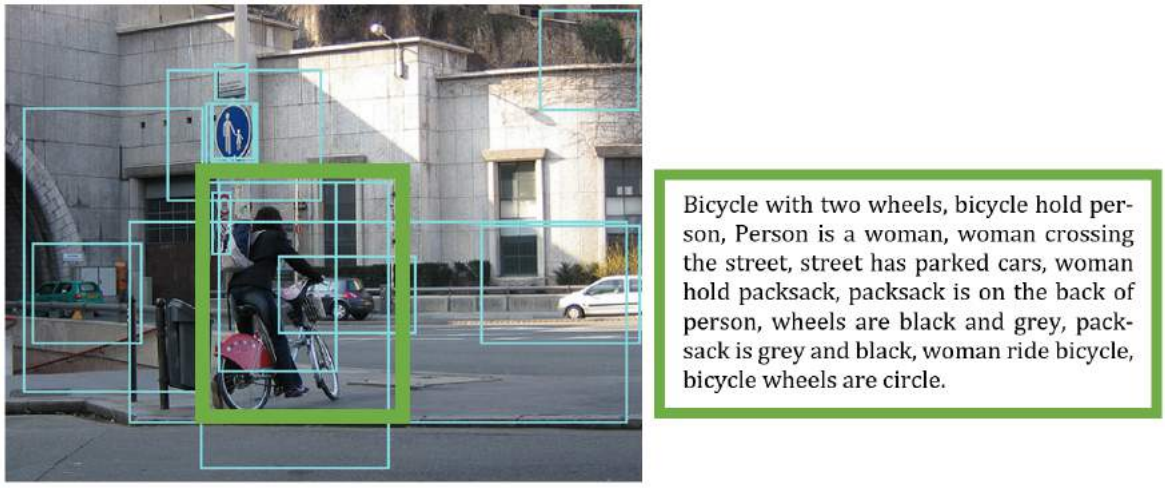


Figure 4.9: An example of manual annotations of an image (it is captured from Genome dataset where it has no predicate or relationship detection)

To ensure the high accuracy and confident results of the ground-truth for our final annotation, we made a test on 300 images of test data using five types of classes: person, street, cars, phone, and stroller. We evaluate how much is good the manual annotation with the judges' results. We asked 6 people to judge the manual annotation, we call them "judges". For each annotation, we confirmed its correctness if the majorities agreed up on it and chose the assignments that corresponds to our manual annotation. The obtained results are displayed in Table 4.1. We evaluated the manual annotation in false (and, true) negative (and positive) using recall, accuracy, and precision.

The low review for the class "stroller" shows that there were uncorrelated examples in different information sources that we took advantage of, or another explanation is seeming the issue of the long tail issue. Hardly any model against tremendous ones. The " stroller" class has additionally the least accuracy and was generally mistaken for other class' explanation.

The judgement list of choices are not only limited to the five classes, judges may also vote with "out of range" for annotations that seems to be irrelevant to total assignment's choices.

Class	Person	Street	Cars	Phone	Stroller	
Recall	93,64%	86,00%	91,67%	82,60%	76,19%	
Precision	91,15%	87,75%	89,80%	90,50%	84,21%	
Accuracy	91,66%	89,28%	90,56%	92,00%	87,50%	
# Samples	110	50	96	23	21	300

Table 4.1: Examination of the consequently created names with the explanations of the five volunteers and the subsequent number of tests per class in the test set

If the majority vote with "out of range" for a certain annotation, it should be removed, and the final test data will be updated. We came to delete 32 annotations from different images.

The corresponding confusion matrix is depicted in Table 4.2. Now, with the use of the final update of test data, 800 images are tested. Annotations are ranked and assigned to the appropriate class that corresponds to a high value of rank. The final results are compared with the initial annotation to tell whether the proposed method would be efficient or not. In the following, We test and compare each proposed module with the basic method without applying any modifications.

Class	Person	Street	Cars	Phone	Stroller	# Samples
Person	103	2	3	1	1	110
Street	3	43	3	0	1	50
Cars	3	4	88	0	1	96
Phone	3	0	1	19	0	23
Stroller	1	0	3	1	16	21

Table 4.2: The corresponding confusion matrix of the test that is made by the judges

4.7.2.2 Statistical Ontology Module Evaluation

To illustrate the efficiency of our proposed methods, we first evaluate the statistical ontology module by two different scenarios: (1) using only object detection module, (2) using statistical ontology module. Based on the ontology inherited from the strong connection between the concept of WordNet, we calculated the error rate of class assignments, and the results are shown in Figure 4.10, we also demonstrated the accuracy obtained under the two different scenarios in Table 4.3.

Note that, the results are shown in Figure 4.10 and Table 4.3 are proof to how deep and objective the object class assignment is. For example, taking the class "Person" and its sub-classes men, women, baby, old women, old-men.. During the ranking process, a high value is given to "women" rather than "Person", and the accuracy will be improved for the process

	Person	Street	Cars	Phone	Stroller
Accuracy(1)	45,02	32,13	54,24	43,22	44,56
Accuracy(2)	67,43	45,28	67,39	50,28	68,57
Accuracy(3)	94,15	90,63	93,22	93,34	88,01

Table 4.3: The accuracy obtained from object detection with applying/non-applying a false negatives/positives filtering. Accuracy(1): Accuracy of Object detection module with non-applying of statistical ontology module, Accuracy(2): Accuracy of Statistical ontology module, Accuracy(3): Accuracy of the entire system for object detection

with the deep and specific detection, whereas the error rate will get higher for the one with the more general detection (i.e. stroller is ahead class of "baby-stroller"; a high rank is given to baby-stroller rather than the general assignment "stroller"). An illustrative example is demonstrated in the next subsection.

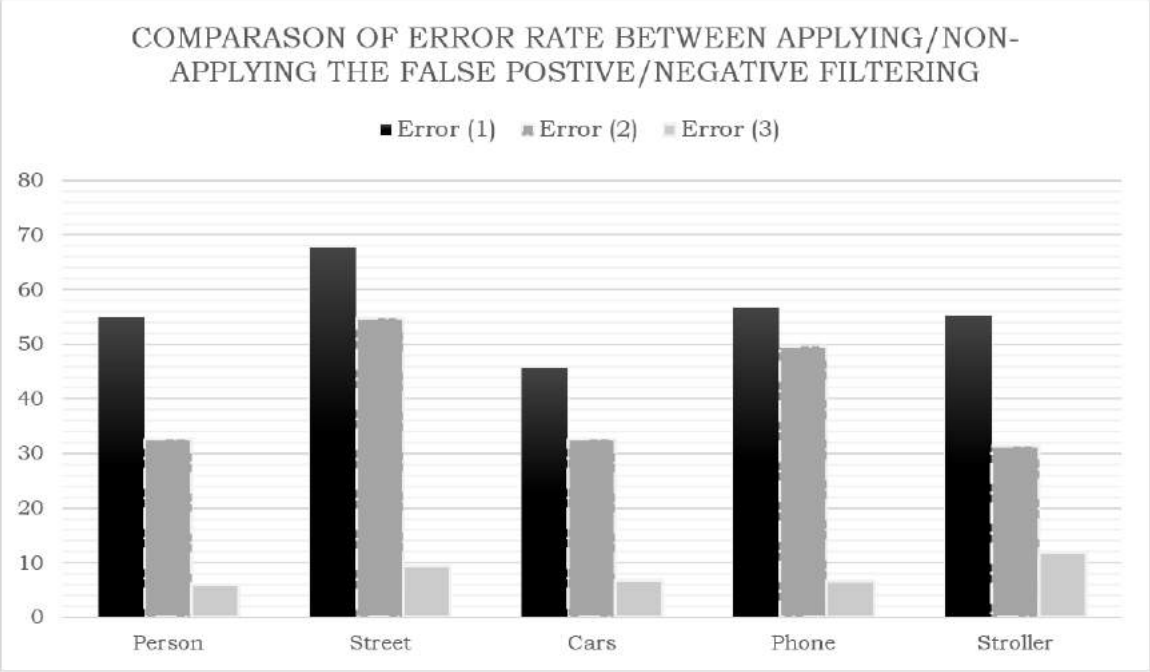


Figure 4.10: The Error rate obtained from object detection of applying/non-applying a false negatives/positives filtering

4.7.2.3 Statistical Ontology Module application

We first demonstrate how the object detection module works with the statistical ontology ranking module, using the example in Figure 4.11. In the object detection phase, (i.e., Visual Genome classification) the class of the boxed object is labeled as "person". And by combining the two modules, a high rank is given to "a tennis-player". The results are due to using the

ontology constraints (i.e., the use of $NC_{value}(o_i)$ and $CA(pr_k^0, k)$) that give a strict and deep assignment of objects. The inappropriate class proposals are filtered before the classification phase. Those that are not semantically close to the context of the image, are penalized. i.e., the context of the image in Figure 4.11. is with the concepts tennis-racquet, a visor a high rank should be assigned to a tennis player rather than a person. The gain of the accuracy that was obtained from this method is presented in Table 4.4.

In Figure 4.12, In the object detection phase, (i.e., Visual Genome classification) the red and the blue boxed objects are labeled as "*person*". And by combining the two modules, a high rank is given to $soccer - player$. Also, a yellow boxed object is labeled as "ball", whereas the proposed statistical ontology ranked it as "$soccer - ball$". The proposed modules in the semantic ontologies helped us to improve the quality of object class assignments by transferring a deeply rich background knowledge. The gain of the accuracy that was obtained from this method is presented in Table 4.5.

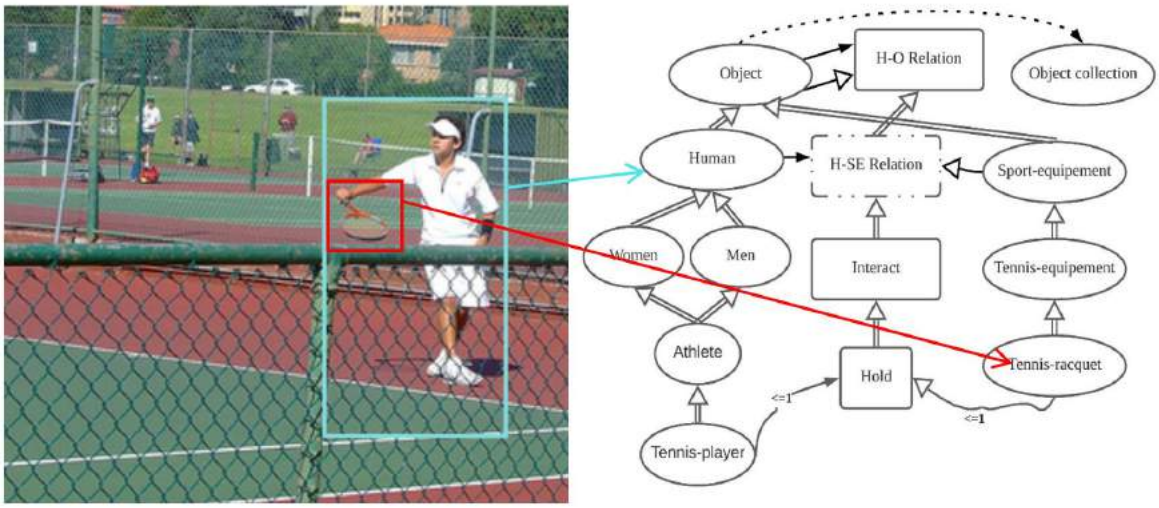


Figure 4.11: An example of a generated ontology from using the detected objects in an image. The explored data of Genome (1) labeled the image as "Racquet", while the proposed statistical ontology module labeled it as "Tennis-player" and "Tennis-racquet"

We change the quantity of tests N for each preparation model, for example 15, 50, 100, 200, 600, 800, 1000, 1200, 1400, 1600, and 1800, to concentrate on the exactness bend of the three analyzed techniques utilizing five kinds of items class, i.e. person, street, cars, phone, and stroller. For one accuracy curve, we kept the same parameters used in session 4.7.2. The obtained results are depicted in Figure 4.13. We had also measured the error rate per number of samples for each training example. The obtained results are depicted in Figure 4.14.

The outcomes affirm our investigation that despite the fact that CBIR shows magnificent execution in object recognition based picture recovery, understanding the semantic of pictures

	$NC_{value}(o_i)$	$CA(pr_k^0, k)$	$S(onto, o_i)$
Person	0.0423	0.0612	0.0974
Men	0.0621	0.0510	0.0734
Women	0.0857	0.0752	0.0913
Athlete	0.3463	0.4043	0.4135
Tennis-player	0.3921	0.4122	0.5194

Table 4.4: Evaluation results of the ranking functions in term of accuracy gain while using the image and the ontology presented in 4.11. The gain in accuracy is by using only $NC_{value}(o_i)$, in the first column, both $NC_{value}(o_i)$ and $CA(pr_k^0, k)$ in the second column, and $NC_{value}(o_i)$, $CA(pr_k^0, k)$ and $S(onto, o_i)$ in the last column

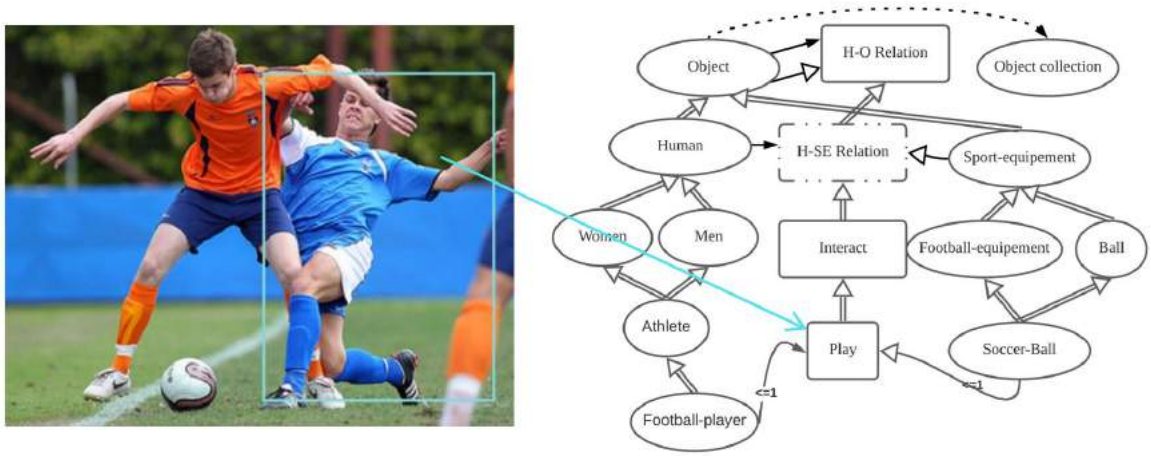


Figure 4.12: The generated ontology based on the detected objects in the image. The explored data of Genome labeled it as "human is playing" and as "a ball", while the use of the statistical ontology module with the advantage of the rich background of the ontology, labeled it as two "football-player" and "Soccer-Ball"

	$NC_{value}(o_i)$	$CA(pr_k^0, k)$	$S(onto, o_i)$
Person	0.0423	0.0486	0.0813
Football-Player1	0.0621	0.0644	0.0713
Football-Player2	0.1057	0.1134	0.2213
Ball	0.0393	0.0453	0.0246
Soccer-Ball	0.2866	0.2964	0.3913

Table 4.5: Evaluation results of the ranking functions in term of accuracy gain while using the image and the ontology presented in Figure 4.12. The gain in accuracy is by using only $NC_{value}(o_i)$, in the first column, both $NC_{value}(o_i)$ and $CA(pr_k^0, k)$ in the second column, and $NC_{value}(o_i)$, $CA(pr_k^0, k)$ and $S(onto, o_i)$ in the last column

with just traditional highlights isn't adequately adequate to conquer the semantic hole and the intra/between class dissimilarity between low-level visual elements and undeniable level

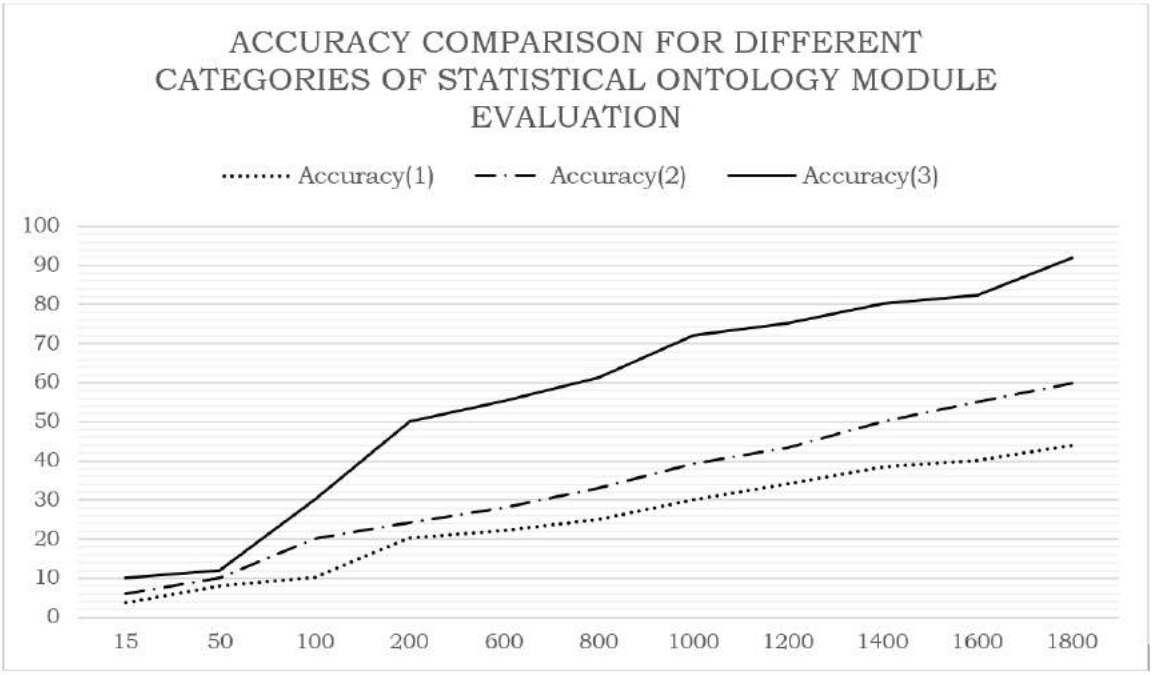


Figure 4.13: Accuracy comparison for different categories of statistical ontology module evaluation

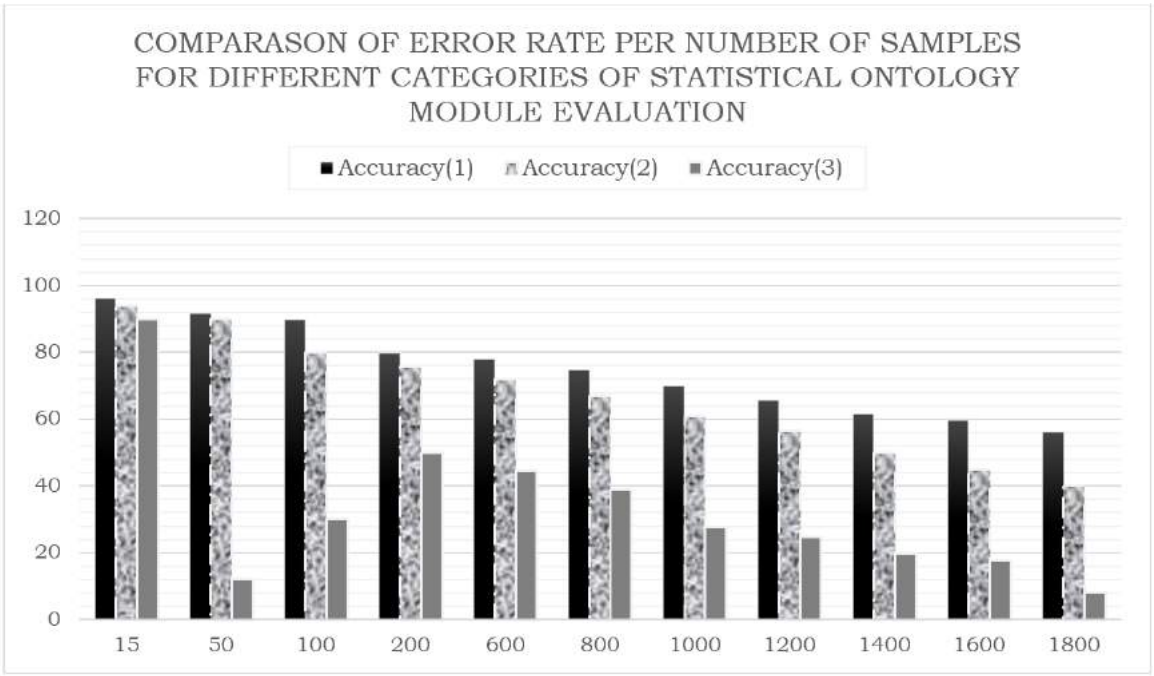


Figure 4.14: Comparison of error rate per number of samples for different categories of statistical ontology module evaluation

semantic element. Furthermore, the key metric that we have used in our work is the use of prior knowledge about how humans interpret the content of image with the aid of ontologies

specialization and specification.

4.7.2.4 Semantic relationship-HO module evaluation

Then, we evaluate the *semantic relationship − HO* ranking module and the visual relationship ranking module by two other different scenarios: (1) using only visual relationship detection module (without high dimension spatial transformation), (2) using *semantic relationship − HO* ranking module (with high dimension spatial transformation). Based on the ontology inherited from the strong connection between the concept of WordNet, we calculated the error rate of class assignments, and the results are shown in Figure 4.15. we also demonstrated the accuracy obtained under the two different scenarios in Table 4.6.

From the obtained results in Table 4.3 and Table 4.6, we can certainly answer the question Q1 presented in Section. 7. The semantic relationship between concepts of ontologies helps in improving the quality of relationship/object class assignment by transferring a deep and rich background knowledge. Also, the obtained results in Figure 4.15 and Table 4.6 are proof of how deep and objective the relationship class prediction is; a high rank is assigned to a specific sub-class rather than the general head-class. An illustrative example is demonstrated in the next sub-section.

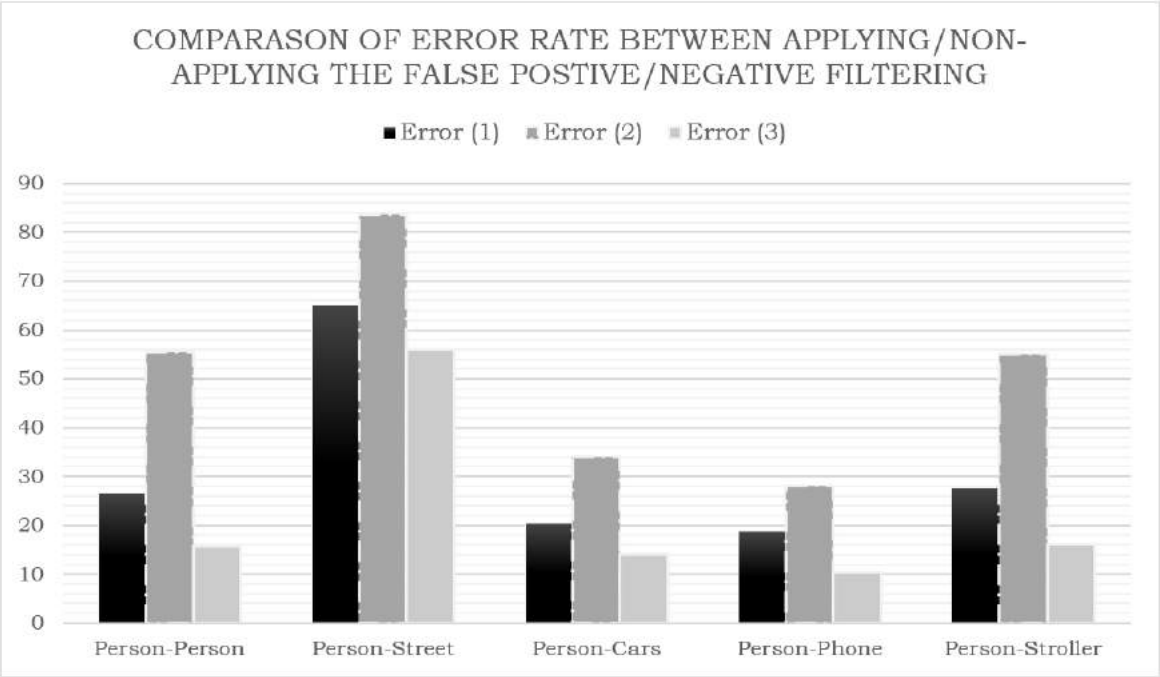


Figure 4.15: The Error rate obtained from the visual relationship detection module (Error(1)) and the semantic relationship-HO ranking module (Error(2)), and the error rate of the entire system Ξ_{onto} (Error(3))

	P-Person	P-Street	P-Cars	P-Phone	P-Stroller
Accuracy (1)	73,43	34,93	79,58	81,22	72,34
Accuracy (2)	44,56	16,38	65,85	71,93	44,95
Accuracy (3)	84,24	63,95	85,84	89,58	83,94

Table 4.6: The accuracy obtained from the visual relationship detection module and the semantic relationship-HO ranking module. P-P: Person-Person, P-S: Person-Street; P-C: Person-Cars; P-Ph: Person-Phone; P-St: Person-Stroller Accuracy(1): is for the semantic relationship-HO ranking module, Accuracy(2): is for the statistical ontology module Accuracy(3): is for the entire system

4.7.2.5 Semantic relationship-HO module Application

In this session, we demonstrate how the Semantic relationship-HO module works with the visual relationship module, using the example in Figure 4.16. In this session, we demonstrate how the semantic relationship-HO module works with the visual relationship module, using the example in Figure 4.16. In the Visual Genome relationship classification, the blue boxed relationship is labeled as "person is playing". Also, other relationships are detected (not illustrated in the figure) such as "shoes on the person", "hand-on the person", "head-on-person"..etc. And by combining the two modules, a high rank is given to "football-player1 is playing with a soccer-ball" "football player1 is interacting with footballer-player2" "football-player1 kick a soccer-ball". The accuracy gain of this method is presented in Table 4.7.

From aside, strong results are obtained from the statistical ontology module that helped in filtering the inappropriate object class proposals. And from another side, they are also obtained because the semantic relationship-HO module penalized the inappropriate human-objects relationship and guarantee deep and strict assignments. these gained advantages are the results of the proposed semantic ontology that helps in improving the quality of relationship assignment by transferring a rich background knowledge.

	Ξ_{onto}	Ξ^{V}_{ops}	Ξ_{total}
Person	0.1239	0.0352	0.1341
Football-player1	0.2425	0.1425	0.2476
Football-Player2	0.3914	0.1355	0.3945
Soccer-Ball	0.1563	0.0313	0.1874
Play	0.3925	0.1827	0.3983
Kick	0.3636	0.1484	0.3713

Table 4.7: Evaluation results of the ranking functions in terms of accuracy gain while using the image and the ontology presented in Figure 4.16. The gain in accuracy is by using only Ξ_{onto} in the first column, Ξ^{V}_{ops} in the second column, and Ξ_{total} in the last column

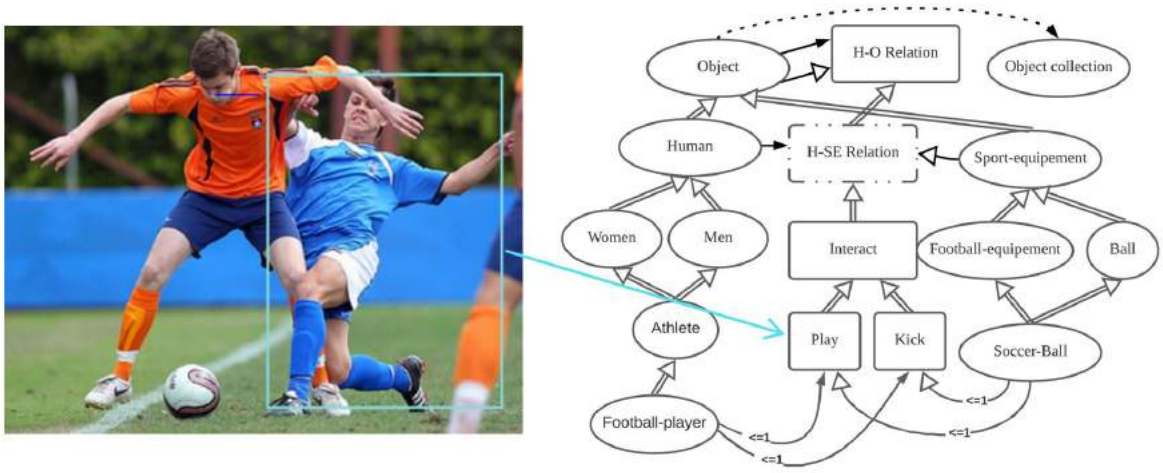


Figure 4.16: An example of a generated ontology from using the detected objects in an image. The explored data of Genome labeled the image as "human is playing" and as "a ball", while the use of the statistical ontology Module, semantic relationship-HO module, and the advantage of the rich background of the ontology, we have been to label it as two "football-players are playing with a soccer-ball"

The same measure as for session 4.7.2.3. we used five types of relationship detection between objects, i.e., person-person, person-street, person-cars, person-phone, person-stroller. We change the quantity of tests N for each preparation model, i.e. 15, 50, 100, 200, 600, 800, 1000, 1200, 1400, 1600, and 1800, to study the accuracy curve of the three compared methods. The accuracy of applying only the visual relationship detection module (Accuracy (1)), the accuracy of applying only the semantic relationship-HO ranking module (Accuracy (2)), and the accuracy of the entire system Ξ_{onto} (Accuracy (3))). For one accuracy curve, we kept the same parameters used in session 4.7.2. The obtained results are depicted in Figure 4.17.

We had also measured the error rate per number of samples for each training example. The error rate of applying only the visual relationship detection module (Error (1)), the error rate of applying only the semantic relationship-HO ranking module (Error (2)), and the error rate of the entire system Ξ_{onto} (Error (3))). The obtained results are depicted in Figure 4.18.

Now to compare our proposed methods, we evaluated under the same assumptions and using the same number of samples in both training set and dataset. The accuracy of the work in (334) is the closest to our work besides the others. This may explain the efficiency of addressing the semantic issues. The results obtained are depicted in Table 4.8. The reason behind the short comparison is that most of the related researches work on object detection and try to enhance the semantic view. Few works addressed the relationship between objects in images or video.

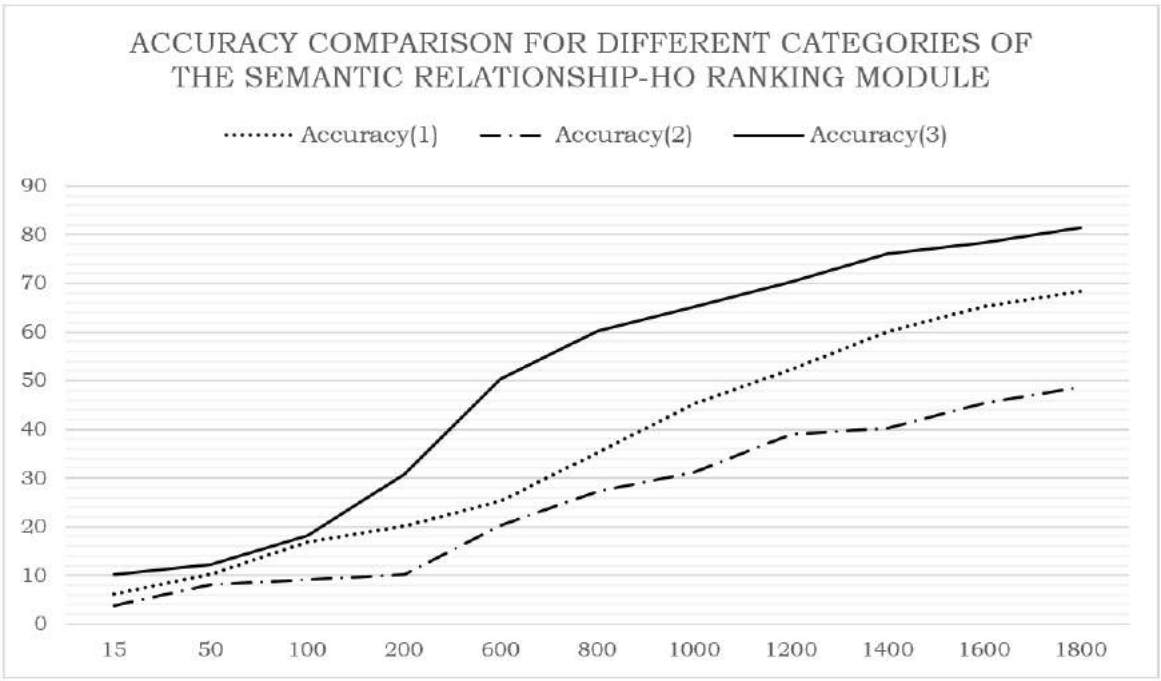


Figure 4.17: Accuracy comparison for different categories of the semantic relationship-HO ranking module evaluation

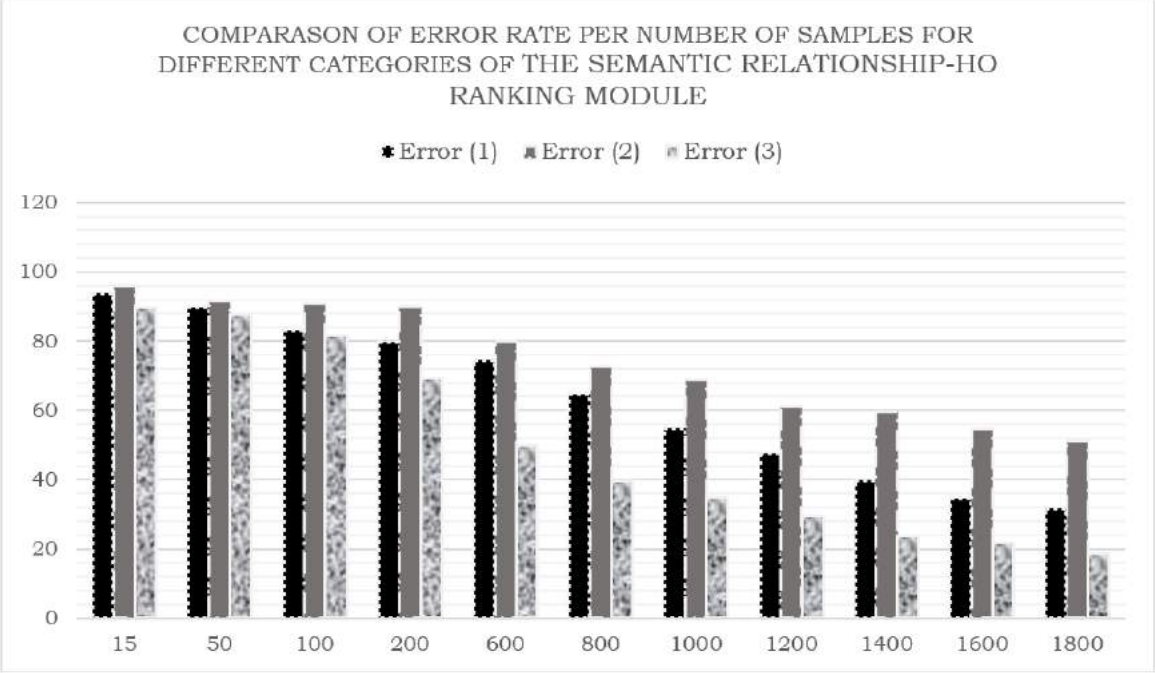


Figure 4.18: Comparison of error rate per number of samples for different categories of the semantic relationship-HO ranking module evaluation

Method	Recall	Accuracy
(335)	-	55%
(334)	-	80,8%
(336)	-	68.8%
Our	85.83	82.39%

Table 4.8: Evaluation results of the ranking functions in terms of accuracy gain while using the image and the ontology presented in Figure 4.16. The gain in accuracy is by using only Ξ_{onto} in the first column, Ξ^{V}_{ops} in the second column, and Ξ_{total} in the last column

4.8 Conclusion and perspectives

In this chapter, we have solved two main problems of relationships detection (i.e., large intra-class divergence and the semantic dependency or semantic gap), and use HCVRD dataset that already solved the third problem (i.e.long-tail problem). We have proposed an ontological semantic model to filter false negatives/positives using a statistical ranking module. Taking the advantage of the ontology rich and strong background knowledge, strong results were obtained from the statistical ontology module that helped in filtering the inappropriate object class proposals. And also from the semantic relationship-HO module that penalized the inappropriate human-objects relationship and guarantees deep and strict assignments. It was offered by the semantic ontology that helps in improving the quality of relationship assignments. As perspectives, we aim in future work to propose the use of the same ontological model and develop in such a way that, a model can tell if 1) pedestrians are crossing the street, 2) how much they are far from a driver, and 3) is it safe or not to keep driving under some condition. The whole perspectives describe the safety of self-driving cars.

5

Summary

This thesis is dedicated to the semantic extraction and interpretation of content-based image retrieval. The main goal is to provide approaches that exploit the semantic contextual information as well as the inter-concept relationship. This is done by defining the context of each object belonging to the image in question. We develop an automatic semantic interpreting system that relies on human semantic interpretation. In our work, we advantage from the advances in philosophy. Ontologies are "an unequivocal determination of a conceptualization". They ensure a common perspective of a specific area, just as a conventional model that is managable to solo machine handling.

Our thesis work has several strengths. The first important point is the genericity of our approaches. Indeed, our work is done in such a way as to achieve our objectives while proposing generic approaches, which can be applied to any indexing system for the detection of any target concept. On the other hand, we insisted that our contributions should not be specific to a particular category of concepts. Another important point of this thesis work is the fact of considering several approaches acting at different levels of an indexing system. This would make it possible to compare and find the most suitable level to effectively exploit the context.

In this thesis, we proposed a semantic relationship detection (SRD) model that is divided into shared three sub-modules. The statistical ontology module, semantic relationship-HO ranking module visual relationship ranking module. For an input image, we use object detection module, we define the corresponding ontology that contains background knowledge of the detected objects and their relations. Then, we apply the statistical ranking module that aims at filtering false negatives/positives results. After that, and in parallel way, We extract the the spatial features and transform them onto a high dimension and in parallel way, we use the ontological module to rank the semantic relationship between <human-object> pairs.

For the ontological module, we used four types of background knowledge; Subsumption, Domain/range, Cardinality constraints, and Collection. Background knowledge is used to define the constrains inter/intra object-subject, i.e., a deep rich informative knowledge is used to describe the object it self as well as it relationship with other object in the same image.

In order to achieve the filtered class proposals, we apply two techniques that are used in an end-to-end manner, namely, C/NC ranking and Contrastive analysis ranking. Both techniques are normalized to fit the CBIR systems. For the C/NC value, it tends to find a group of the class proposals that are with high prediction probabilities, and considers the high predicted class proposals, and tries to find the most frequently appeared of those class proposals.

The contrastive analysis is used to eliminate terms that are not relevant to the context. Filtering ensures that the proposals that are more relevant to the context, shall stay and that is done by measuring the specificity of a term with respect to its range.

For the visual relationship ranking module, we started by extracting the corresponding features, we feed the common convolutional include map and the created object proposition to a $ROI - pooling$ layer and we acquired the subject and article visual component vectors. After that, We transform the vectors onto a high dimension geometric spatial location and transform them into features. Then, we rank the visual relationship proposal using the geometric transformation of each bounding box spatial characteristic; a probability of each object is obtained.

The best possible naming is needed to fulfill the metaphysics sayings and limitations, be profoundly and lavishly useful, and expand the position of the class task of every trio $<$ Human$-$object $>$ concerning their visual appearance.

5.1 Perspectives & Future Work

Our model has been developed while considering images of HCVRD, Genome, and VRD datasets. We built background knowledge ontologies using examples of athletes. Trying to build for each object a rich background is time-consuming. For that, the most important thing is to choose the right benchmarks and case studies. As perspectives, intend in future works to use the same model and extend to fit another source of knowledge such as videos. Another interesting research field that needs to invest, is traffic accidents. It would be very good if someone gets a notification about a sudden passing by of a person on a street. Traffic accidents aren't fault-tolerant, building an alarms system before an accident happens

is crucial. This may be done using the rich background of ontologies. There exist a system [1] that detect the shape of the street, but they don't tell us how much a pedestrian is far from a car in real-time. We intend in future work to investigate the safety of self-driving cars.

In VRD dataset, predicates are categorized into five types: Action, Spatial, Preposition, Comparative, and Verb. Since ontologies building are time consuming, we took much time in our model to carry only two types of predicates; action and spatial. In future work and as extension work, we will apply the same techniques on the remain type of predicates, i.e. comparative, verb, and preposition.

[1] https://www.utoronto.ca/news/these-u-t-experts-ai-want-teach-you-how-program-self-driving-car

References

[1] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. **Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations**. 2016.

[2] Jun Shu, Zongben Xu, and Deyu Meng. **Small sample learning in big data era**. *arXiv preprint arXiv:1808.04572*, 2018.

[3] Li Zhou, Jian Zhao, Jianshu Li, Li Yuan, and Jiashi Feng. **Object Relation Detection Based on One-shot Learning**. *arXiv preprint arXiv:1807.05857*, 2018.

[4] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. **Hcvrd: a benchmark for large-scale human-centered visual relationship detection**. 2018.

[5] Jonathon S Hare, Paul H Lewis, Peter GB Enser, and Christine J Sandom. **Mind the Gap: Another look at the problem of the semantic gap in image retrieval**. *Multimedia Content Analysis, Management, and Retrieval 2006*, **6073**:607309, 2006.

[6] W. Zhou, H. Li, and Q. Tian. **Recent Advance in Content-based Image Retrieval: A Literature Survey**. *ArXiv*, **abs/1706.06064**, 2017.

[7] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. M. Snoek, and A. Del Bimbo. **Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement and Retrieval**. *ACM Computing Surveys*, 2016.

[8] Ahmad Alzu'bi, Abbes Amira, and Naeem Ramzan. **Semantic content-based image retrieval: A comprehensive study**. *Journal of Visual Communication and Image Representation*, **32**:20–54, October 2015.

[9] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. **Content-based image retrieval at the end of the early years**. *IEEE Transactions on pattern analysis and machine intelligence*, **22**(12):1349–1380, 2000.

[10] S-F Cheng, William Chen, and Hari Sundaram. **Semantic visual templates: linking visual features to semantics**. *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)*, pages 531–535, 1998.

[11] Yueting Zhuang, Xiaoming Liu, and Yunhe Pan. **Apply semantic template to support content-based image retrieval**. *Storage and Retrieval for Media Databases 2000*, **3972**:442–449, 1999.

[12] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. **A survey of content-based image retrieval with high-level semantics**. *Pattern recognition*, **40**(1):262–282, 2007.

[13] Nancy Goyal and Navdeep Singh. **A review on different content based image retrieval techniques using high level semantic features**. *International Journal of Innovative Research in Computer and Communication Engineering*, **2**(7), 2014.

[14] Brigit Schroeder and Subarna Tripathi. **Structured Query-Based Image Retrieval Using Scene Graphs**. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[15] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Barbara Shaw, Wessel Kraaij, Alan F Smeaton, and Georges Quéot. **Trecvid 2012-an overview of the goals, tasks, data, evaluation mechanisms and metrics**. 2013.

[16] Manish Bhattarai, D. Oyen, Juan Castorena, L. Yang, and B. Wohlberg. **Diagram Image Retrieval using Sketch-Based Deep Learning and Transfer Learning**. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 663–672, 2020.

[17] Atsuo Yoshitaka and Tadao Ichikawa. **A survey on content-based retrieval for multimedia databases**. *IEEE Transactions on Knowledge and Data Engineering*, **11**(1):81–93, 1999.

[18] Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. **Content-based multimedia information retrieval: State of the art and challenges**. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, **2**(1):1–19, 2006.

[19] Bhagwandas Patel, Kuldeep Yadav, and Debashis Ghosh. **Current Trend and Methodologies of Content-Based Image Retrieval: Survey**. *Proceedings of Second International Conference on Smart Energy and Communication*, pages 647–665, 2021.

[20] Shiv Ram Dubey. **A Decade Survey of Content Based Image Retrieval using Deep Learning**. *arXiv preprint arXiv:2012.00641*, 2020.

[21] Aleksandra Mojsilovic and Bernice Rogowitz. **Capturing image semantics with low-level descriptors**. *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, **1**:18–21, 2001.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. **Imagenet classification with deep convolutional neural networks**. *Advances in neural information processing systems*, **25**:1097–1105, 2012.

[23] Ishwar K Sethi, Ioana L Coman, and Daniela Stan. **Mining association rules between low-level image features and high-level concepts**. *Data mining and knowledge discovery: theory, tools, and technology III*, **4384**:279–290, 2001.

[24] Najlae Idrissi, José Martinez, and Driss Aboutajdine. **Bridging the semantic gap for texture-based image retrieval and navigation.** *Journal of Multimedia*, **4**(5), 2009.

[25] Thomas R Gruber. **A translation approach to portable ontology specifications**. *Knowledge acquisition*, **5**(2):199–220, 1993.

[26] Zhiliang Ma, Zhe Liu, and Zhenhua Wei. **Formalized representation of specifications for construction cost estimation by using ontology**. *Computer-Aided Civil and Infrastructure Engineering*, **31**(1):4–17, 2016.

[27] Dana H Ballard and Christopher M Brown. **Computer vision. englewood cliffs**. *J: Prentice Hall*, 1982.

[28] Thomas Huang. **Computer vision: Evolution and promise**. 1996.

[29] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[30] Reinhard Klette. *Concise computer vision*. Springer, 2014.

[31] Linda G Shapiro and G Stockman. **Computer vision prentice hall**. *Inc., New Jersey*, 2001.

[32] Tim Morris. *Computer vision and image processing*. Palgrave Macmillan Ltd, 2004.

[33] Bernd Jähne and Horst Haussecker. **Computer vision and applications**. 2000.

[34] David A Forsyth and Jean Ponce. *Computer vision: a modern approach*. Pearson,, 2012.

[35] Luis Von Ahn and Laura Dabbish. **Labeling images with a computer game**. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, 2004.

[36] Haihua Feng, David A Castanon, and William Clement Karl. **A curve evolution approach for image segmentation using adaptive flows**. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, **2**:494–499, 2001.

[37] Koen EA Van de Sande, Theo Gevers, and Cees GM Snoek. **A comparison of color features for visual concept classification**. *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 141–150, 2008.

[38] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. **Textural features corresponding to visual perception**. *IEEE Transactions on Systems, man, and cybernetics*, **8**(6):460–473, 1978.

[39] Fang Liu and Rosalind W Picard. **Periodicity, directionality, and randomness: Wold features for image modeling and retrieval**. *IEEE transactions on pattern analysis and machine intelligence*, **18**(7):722–733, 1996.

[40] Bangalore S Manjunath and Wei-Ying Ma. **Texture features for browsing and retrieval of image data**. *IEEE Transactions on pattern analysis and machine intelligence*, **18**(8):837–842, 1996.

[41] Lance M Kaplan, Romain Murenzi, and Kameswara Rao Namuduri. **Fast texture database retrieval using extended fractal features**. *Storage and retrieval for image and video databases VI*, **3312**:162–173, 1997.

[42] Farzin Mokhtarian, Sadegh Abbasi, and Josef Kittler. **Efficient and robust retrieval by shape content through curvature scale space**. pages 51–58. World Scientific, 1997.

[43] P Bonton, C Fernandez-Maloigne, and A Tremeau. **IMAGE NUMERIQUE COULEUR De l'acquisition au traitement**. *DUNOD, ISBN*, **2**(10):006843, 2004.

[44] B Srinivasa Reddy and Biswanath N Chatterji. **An FFT-based technique for translation, rotation, and scale-invariant image registration**. *IEEE transactions on image processing*, **5**(8):1266–1271, 1996.

[45] Ming-Kuei Hu. **Visual pattern recognition by moment invariants**. *IRE transactions on information theory*, **8**(2):179–187, 1962.

[46] David G Lowe. **Distinctive image features from scale-invariant keypoints**. *International journal of computer vision*, **60**(2):91–110, 2004.

[47] Christian Jutten and Jeanny Herault. **Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture**. *Signal processing*, **24**(1):1–10, 1991.

[48] Aapo Hyvärinen and Erkki Oja. **Independent component analysis: algorithms and applications**. *Neural networks*, **13**(4-5):411–430, 2000.

[49] Mauro Dalla Mura, Alberto Villa, Jon Atli Benediktsson, Jocelyn Chanussot, and Lorenzo Bruzzone. **Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis**. *IEEE Geoscience and Remote Sensing Letters*, **8**(3):542–546, 2010.

[50] Jing Wang and Chein-I Chang. **Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis**. *IEEE transactions on geoscience and remote sensing*, **44**(6):1586–1600, 2006.

[51] Wei Xia, Xuesong Liu, Bin Wang, and Liming Zhang. **Independent component analysis for blind unmixing of hyperspectral imagery with additional constraints**. *IEEE transactions on geoscience and remote sensing*, **49**(6):2165–2179, 2011.

[52] KC Tiwari, Manoj K Arora, and Dharmendra Singh. **An assessment of independent component analysis for detection of military targets from hyperspectral images**. *International Journal of Applied Earth Observation and Geoinformation*, **13**(5):730–740, 2011.

[53] Mi-Hye Song, Jeon Lee, Sung-Pil Cho, Kyoung-Joung Lee, and Sun-Kook Yoo. **Support vector machine based arrhythmia classification using reduced features**. *International Journal of Control, Automation, and Systems*, **3**(4):571–579, 2005.

[54] Wei-Ying Ma and Bangalore S Manjunath. **Netra: A toolbox for navigating large image databases**. *Multimedia systems*, **7**(3):184–198, 1999.

[55] Yuqing Song, Wei Wang, and Aidong Zhang. **Automatic annotation and retrieval of images**. *World Wide Web*, **6**(2):209–231, 2003.

[56] Aleksandra Mojsilovic, José Gomes, and Bernice E Rogowitz. **Isee: Perceptual features for image library navigation**. *Human Vision and Electronic Imaging VII*, **4662**:266–277, 2002.

[57] Laurent Guigues, Roger Trias-Sanz, Nesrine Chehata, Franck Taillandier, and Matthieu Deveau. **B. 4 Segmentation multi-échelles d'images: théorie et applications**. *Bulletin d'Information Scientifique & Technique de l'IGN n*, **75**(1):41, 2006.

[58] Wolfgang Förstner. **A framework for low level feature extraction**. *European Conference on Computer Vision*, pages 383–394, 1994.

[59] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. **Image retrieval: Ideas, influences, and trends of the new age**. *ACM Computing Surveys (Csur)*, **40**(2):1–60, 2008.

[60] Amarnath Gupta and Ramesh Jain. **Visual information retrieval**. *Communications of the ACM*, **40**(5):70–79, 1997.

[61] Yong Rui, Thomas S. Huang, and Shih-Fu Chang. **Image Retrieval: Current Techniques, Promising Directions, and Open Issues**. *Journal of Visual Communication and Image Representation*, **10**(1):39–62, 1999.

[62] Jiebo Luo, A. Savakis, and A. Singhal. **A Bayesian network-based framework for semantic image understanding**. *Pattern Recognit.*, **38**:919–934, 2005.

REFERENCES

[63] Arnab Ghoshal, Pavel Ircing, and Sanjeev Khudanpur. **Hidden Markov models for automatic annotation and content-based retrieval of images and video**. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 544–551, 2005.

[64] Tao Wang, Yong Rui, Shi-min Hu, and Jia-guang Sun. **Adaptive tree similarity learning for image retrieval**. *Multimedia Systems*, **9**(2):131–143, 2003.

[65] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. **Learning deep representations of fine-grained visual descriptions**. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–58, 2016.

[66] Yong Rui, Thomas S Huang, and Shih-Fu Chang. **Image retrieval: Current techniques, promising directions, and open issues**. *Journal of visual communication and image representation*, **10**(1):39–62, 1999.

[67] Carlo Meghini, Fabrizio Sebastiani, and Umberto Straccia. **A model of multimedia information retrieval**. *Journal of the ACM (JACM)*, **48**(5):909–970, 2001.

[68] Huan Wang, Song Liu, and Liang-Tien Chia. **Does ontology help in image retrieval? A comparison between keyword, text ontology and multi-modality ontology approaches**. *Proceedings of the 14th ACM international conference on multimedia*, pages 109–112, 2006.

[69] Michele Trevisiol, Luca Chiarandini, Luca Maria Aiello, and Alejandro Jaimes. **Image ranking based on user browsing behavior**. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 445–454, 2012.

[70] Bin Xu, Jiajun Bu, Chun Chen, Deng Cai, Xiaofei He, Wei Liu, and Jiebo Luo. **Efficient manifold ranking for image retrieval**. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 525–534, 2011.

[71] Y Alp Aslandogan and Clement T. Yu. **Techniques and systems for image and video retrieval**. *IEEE transactions on Knowledge and Data Engineering*, **11**(1):56–63, 1999.

[72] Mohamed Daoudi and Stanislaw Matusiak. **Visual image retrieval by multiscale description of user sketches**. *Journal of Visual Languages & Computing*, **11**(3):287–301, 2000.

[73] Eugenio Di Sciascio, G Mingolla, and Marina Mongiello. **Content-based image retrieval over the web using query by sketch and relevance feedback**. *International Conference on Advances in Visual Information Systems*, pages 123–130, 1999.

[74] Romain Negrel. *Représentations optimales pour la recherche dans les bases d'images patrimoniales*. PhD thesis, Cergy-Pontoise, 2014.

[75] Allan Hanbury. **A survey of methods for image annotation**. *Journal of Visual Languages & Computing*, **19**(5):617–627, 2008.

[76] Jean Charlet, Bruno Bachimont, and Raphaël Troncy. **Ontologies pour le web sémantique**. *Revue I3, numéro Hors Série «Web sémantique*, pages 43–63, 2004.

[77] Jérôme Champavère. **De la représentation des connaissances au web sémantique: Un survol**. *Tiré de: http://www. grappa. univlille3. fr/champavere/Enseignement/0809/l3miashs/ia/rc-ws. pdf*, 2013.

[78] Philippe Aigrain, HongJiang Zhang, and Dragutin Petkovic. **Content-based representation and retrieval of visual media: A state-of-the-art review**. *Multimedia tools and applications*, **3**(3):179–202, 1996.

[79] Jang-Jong Fan and Keh-Yih Su. **An efficient algorithm for matching multiple patterns**. *IEEE Transactions on Knowledge and Data Engineering*, **5**(2):339–351, 1993.

[80] Gobinda G Chowdhury. *Introduction to modern information retrieval*. Facet publishing, 2010.

[81] Kosmas Petridis, Dionysios Anastasopoulos, Carsten Saathoff, Norman Timmermann, Yiannis Kompatsiaris, and Steffen Staab. **M-ontomat-annotizer: Image annotation linking ontologies and multimedia low-level features**. *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 633–640, 2006.

[82] Anil K Jain and Aditya Vailaya. **Image retrieval using color and shape**. *Pattern recognition*, **29**(8):1233–1244, 1996.

[83] Maristella Agosti, Giorgetta Bonfiglio-Dosio, and Nicola Ferro. **A historical and contemporary study on annotations to derive key features for systems design**. *International Journal on Digital Libraries*, **8**(1):1–19, 2007.

[84] Stéphane Bres, Jean-Michel Jolion, and Frank Lebourgeois. *Traitement et analyse des images numériques*. Hermes Science Publications, 2003.

[85] Diane Lingrand. *Introduction au Traitement d'images*. Vuibert, 2008.

[86] John Douglas Bradley. **Pliny: A model for digital support of scholarship**. *Journal of Digital Information*, **9**(1), 2008.

[87] Gustavo Carneiro, Antoni B Chan, Pedro J Moreno, and Nuno Vasconcelos. **Supervised learning of semantic classes for image annotation and retrieval**. *IEEE transactions on pattern analysis and machine intelligence*, **29**(3):394–410, 2007.

[88] Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha. **Automatic image annotation and retrieval using cross-media relevance models**. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126, 2003.

[89] Rong Jin, Joyce Y Chai, and Luo Si. **Effective automatic image annotation via a coherent language model and active learning**. *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 892–899, 2004.

[90] Yul Gao and Jianping Fan. **Incorporating concept ontology to enable probabilistic concept reasoning for multi-level image annotation**. *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 79–88, 2006.

[91] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. **Baselines for image annotation**. *International Journal of Computer Vision*, **90**(1):88–105, 2010.

[92] Christophe Millet. *Annotation automatique d'images: annotation cohérente et création automatique d'une base d'apprentissage*. PhD thesis, PhD thesis, 2008, ENST, Paris, 2008.

[93] Liu Wenyin, Susan T Dumais, Yanfeng Sun, HongJiang Zhang, Mary Czerwinski, Brent A Field, et al. **Semi-Automatic Image Annotation.** *Interact*, **1**:326–333, 2001.

[94] Roger CF Wong and Clement HC Leung. **Automatic semantic annotation of real-world web images**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**(11):1933–1944, 2008.

[95] Andrew D Bagdanov, Marco Bertini, Alberto Del Bimbo, Giuseppe Serra, and Carlo Torniai. **Semantic annotation and retrieval of video events using multimedia ontologies**. *International Conference on Semantic Computing (ICSC 2007)*, pages 713–720, 2007.

[96] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando De Freitas, David M Blei, and Michael I Jordan. **Matching words and pictures**. 2003.

[97] Stamatia Dasiopoulou and Ioannis Kompatsiaris. **Trends and issues in description logics frameworks for image interpretation**. *Hellenic Conference on Artificial Intelligence*, pages 61–70, 2010.

[98] Jianping Fan, Yuli Gao, and Hangzai Luo. **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, 2008.

[99] Hakim Hacid. **Annotation semi-automatique de grandes BD images: Approche par graphes de voisinage.** *CORIA*, pages 205–211, 2006.

[100] Guangcan Liu, Zhouchen Lin, and Yong Yu. **Radon representation-based feature descriptor for texture classification**. *IEEE Transactions on Image Processing*, **18**(5):921–928, 2009.

[101] Nhu Van Nguyen. *Représentations visuelles de concepts textuels pour la recherche et l'annotation interactives d'images.* PhD thesis, Université de La Rochelle, 2011.

[102] Vassilios Stathopoulos, Jana Urban, and Joemon Jose. **Semantic relationships in multi-modal graphs for automatic image annotation**. *European Conference on Information Retrieval*, pages 490–497, 2008.

[103] Pham Minh Hai. **Apprentissage automatique. travail d'intérêt personnel encadré**. *Institut de la Francophonie pour l Informatique*, 2004.

[104] Christopher JC Burges. **A tutorial on support vector machines for pattern recognition**. *Data mining and knowledge discovery*, **2**(2):121–167, 1998.

[105] Leo Breiman. **Random forests**. *Machine learning*, **45**(1):5–32, 2001.

[106] Sabine Barrat and Salvatore Tabbone. **Classification and automatic annotation extension of images using Bayesian network**. *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 937–946, 2008.

[107] Christopher Town and David Sinclair. **Content based image retrieval using semantic visual categories**, 2000.

[108] Alain Boucher and Thi-Lan Le. **Comment extraire la sémantique d'une image?** *Conference Internationale Sciences Electroniques, Technologies de l'Information et des Telecommunications (SETIT'05)*, pages 295–306, 2005.

[109] Shao Lei Feng, Raghavan Manmatha, and Victor Lavrenko. **Multiple bernoulli relevance models for image and video annotation**. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, **2**:II–II, 2004.

[110] David M Blei and Michael I Jordan. **Modeling annotated data**. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, 2003.

[111] Ruofei Zhang, Zhongfei Zhang, Mingjing Li, Wei-Ying Ma, and Hong-Jiang Zhang. **A probabilistic semantic model for image annotation and multimodal image retrieval**. *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, **1**:846–851, 2005.

[112] Salvatore TABBONE et al. **Classification et extension automatique d'annotations d'images en utilisant un réseau Bayésien**. *Traitement du Signal*, 2009.

[113] Deng Cai, Xiaofei He, Zhiwei Li, Wei-Ying Ma, and Ji-Rong Wen. **Hierarchical clustering of www image search results using visual, textual and link information**. *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 952–959, 2004.

[114] Huamin Feng, Rui Shi, and Tat-Seng Chua. **A bootstrapping framework for annotating and retrieving WWW images**. *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 960–967, 2004.

[115] Deng Cai, Xiaofei He, Wei-Ying Ma, Ji-Rong Wen, and Hongjiang Zhang. **Organizing WWW images based on the analysis of page layout and web link structure**. *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, **1**:113–116, 2004.

[116] Changhu Wang, Feng Jing, Lei Zhang, and Hong-Jiang Zhang. **Image annotation refinement using random walk with restarts**. *Proceedings of the 14th ACM international conference on Multimedia*, pages 647–650, 2006.

[117] Laura Hollink, Guus Schreiber, Jan Wielemaker, Bob Wielinga, et al. **Semantic annotation of image collections**. *Knowledge capture*, **2**, 2003.

[118] Roberto Bartolini, Emiliano Giovannetti, Simone Marchi, Simonetta Montemagni, Claudio Andreatta, Roberto Brunelli, Rodolfo Stecher, and Paolo Bouquet. **Multimedia Information Extraction in Ontology-based Semantic Annotation of Product Catalogues.** *SWAP*, 2006.

[119] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. **Semantic annotation, indexing, and retrieval**. *Journal of Web Semantics*, **2**(1):49–79, 2004.

[120] Gediminas Adomavicius and Alexander Tuzhilin. **Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions**. *IEEE transactions on knowledge and data engineering*, **17**(6):734–749, 2005.

[121] Isaak Kavasidis, Simone Palazzo, Roberto Di Salvo, Daniela Giordano, and Concetto Spampinato. **An innovative web-based collaborative platform for video annotation**. *Multimedia Tools and Applications*, **70**(1):413–432, 2014.

[122] Johanna Vompras and Stefan Conrad. **A semi-automated Framework for Supporting Semantic Image Annotation.** *SemAnnot@ ISWC*, 2005.

[123] Dominik Renzel, Yiwei Cao, Michael Lottko, and Ralf Klamma. **Collaborative video annotation for multimedia sharing between experts and amateurs**. *Proceedings of the 11th international workshop of the multimedia metadata community on interoperable social multimedia applications (WISMA-2010). vol. CEUR workshop proceedings*, pages 7–14, 2010.

[124] Chih-Fong Tsai, Ken McGarry, and John Tait. **Qualitative evaluation of automatic assignment of keywords to images**. *Information processing & management*, **42**(1):136–154, 2006.

[125] Florent Monay and Daniel Gatica Perez. **On image auto annotation with latent space models**. *Proceedings of the eleventh ACM international conference on Multimedia*, pages 275–278, 2003.

REFERENCES

[126] Ville Viitaniemi and Jorma Laaksonen. **Keyword detection approach to automatic image annotation**. 2005.

[127] Erin L Allwein, Robert E Schapire, and Yoram Singer. **Reducing multiclass to binary: A unifying approach for margin classifiers**. *Journal of machine learning research*, **1**(Dec):113–141, 2000.

[128] Ammar Mahdhaoui. *Analyse de signaux sociaux pour la modélisation de l'interaction face à face*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2010.

[129] Frank Dellaert. **The expectation maximization algorithm**. Technical report, Georgia Institute of Technology, 2002.

[130] J Aldrich and RA Fisher. **the making of maximum likelihood 1912-22**. *Department of Economics, University of Southampton, United Kingdom*, 1995.

[131] Makoto Iwayama and Takenobu Tokunaga. **Hierarchical Bayesian clustering for automatic text classification**. *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2*, pages 1322–1327, 1995.

[132] Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz. **A Bayesian approach to filtering junk e-mail**. *Learning for Text Categorization: Papers from the 1998 workshop*, **62**:98–105, 1998.

[133] Jefferson Provost. **Naıve-bayes vs. rule-learning in classification of email**. *University of Texas at Austin*, 1999.

[134] Vladimir N Vapnik. **An overview of statistical learning theory**. *IEEE transactions on neural networks*, **10**(5):988–999, 1999.

[135] Andrew Y Ng and Michael I Jordan. **On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes**. *Advances in neural information processing systems*, pages 841–848, 2002.

[136] Gregory Shakhnarovich, Trevor Darrell, and Piotr Indyk. **Nearest-neighbor methods in learning and vision**. *Neural Information Processing*, 2005.

[137] Avrim L Blum and Pat Langley. **Selection of relevant features and examples in machine learning**. *Artificial intelligence*, **97**(1-2):245–271, 1997.

[138] Anil K Jain, Robert P. W. Duin, and Jianchang Mao. **Statistical pattern recognition: A review**. *IEEE Transactions on pattern analysis and machine intelligence*, **22**(1):4–37, 2000.

[139] Bernhard Schölkopf and Alex Smola. **Support vector machines and kernel algorithms**. pages 5328–5335. Wiley, 2005.

[140] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. **A training algorithm for optimal margin classifiers**. *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[141] Corinna Cortes and Vladimir Vapnik. **Support-vector networks**. *Machine learning*, **20**(3):273–297, 1995.

[142] Olivier Chapelle, Patrick Haffner, and Vladimir N Vapnik. **Support vector machines for histogram-based image classification**. *IEEE transactions on Neural Networks*, **10**(5):1055–1064, 1999.

[143] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[144] Cheng-Lin Liu, Kazuki Nakashima, Hiroshi Sako, and Hiromichi Fujisawa. **Handwritten digit recognition using state-of-the-art techniques**. *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*, pages 320–325, 2002.

[145] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[146] Harris Drucker, Chris JC Burges, Linda Kaufman, Alex Smola, Vladimir Vapnik, et al. **Support vector regression machines**. *Advances in neural information processing systems*, **9**:155–161, 1997.

[147] Bahjat Safadi and Georges Quénot. **Evaluations of multi-learner approaches for concept indexing in video documents**. pages 88–91. 2010.

[148] Robert E Schapire. **The strength of weak learnability**. *Machine learning*, **5**(2):197–227, 1990.

[149] Yoav Freund and Robert E Schapire. **A decision-theoretic generalization of on-line learning and an application to boosting**. *Journal of computer and system sciences*, **55**(1):119–139, 1997.

[150] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. **Multi-class adaboost**. *Statistics and its Interface*, **2**(3):349–360, 2009.

[151] David H Wolpert. **Stacked generalization**. *Neural networks*, **5**(2):241–259, 1992.

[152] Magdalena Graczyk, Tadeusz Lasota, Bogdan Trawiński, and Krzysztof Trawiński. **Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal**. *Asian conference on intelligent information and database systems*, pages 340–350, 2010.

[153] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. **Deep residual learning for image recognition**. pages 770–778, 2016.

[154] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. **Improving neural networks by preventing co-adaptation of feature detectors**. *arXiv preprint arXiv:1207.0580*, 2012.

[155] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. **Dropout: a simple way to prevent neural networks from overfitting**. *The journal of machine learning research*, **15**(1):1929–1958, 2014.

[156] Connor Shorten and Taghi M Khoshgoftaar. **A survey on image data augmentation for deep learning**. *Journal of Big Data*, **6**(1):1–48, 2019.

[157] Xavier Glorot and Yoshua Bengio. **Understanding the difficulty of training deep feedforward neural networks**. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[158] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. **Delving deep into rectifiers: Surpassing human-level performance on imagenet classification**. *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[159] Geoffrey E Hinton and Ruslan R Salakhutdinov. **Reducing the dimensionality of data with neural networks**. *science*, **313**(5786):504–507, 2006.

[160] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. **Convolutional networks and applications in vision**. *Proceedings of 2010 IEEE international symposium on circuits and systems*, pages 253–256, 2010.

[161] Jürgen Schmidhuber. **Deep learning in neural networks: An overview**. *Neural networks*, **61**:85–117, 2015.

[162] L Deng, D Yu, et al. **Deep Learning: Method and Applications. Foundations and TrendsR in Signal Processing 7: 197–387**, 2014.

[163] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. **Multi-column deep neural networks for image classification**. *2012 IEEE conference on computer vision and pattern recognition*, pages 3642–3649, 2012.

[164] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. **Self-taught clustering**. *Proceedings of the 25th international conference on Machine learning*, pages 200–207, 2008.

[165] Sumit Saha. **A comprehensive guide to convolutional neural networks—the eli5 way, 2018**. *URL https://towardsdatascience. com/a-comprehensive-guideto-convolutional-neural-networks-the-eli5-way-3bd2b1164a53. Accessed August*, 2019.

[166] Sinno Jialin Pan and Qiang Yang. **A Survey on Transfer Learning IEEE Transactions on Knowledge and Data Engineering**. *22 (10): 1345*, **1359**, 2010.

[167] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, **1**. MIT press Cambridge, 2016.

[168] Zheng Wang, Yangqiu Song, and Changshui Zhang. **Transferred dimensionality reduction**. *Joint European conference on machine learning and knowledge discovery in databases*, pages 550–565, 2008.

[169] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. **Deep learning scaling is predictable, empirically**. *arXiv preprint arXiv:1712.00409*, 2017.

[170] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. **Imagenet large scale visual recognition challenge**. *International journal of computer vision*, **115**(3):211–252, 2015.

[171] Jeremy Jordan. **Common architectures in convolutional neural networks**. *Jeremy Jordan*, 2018.

[172] Karen Simonyan and Andrew Zisserman. **Very deep convolutional networks for large-scale image recognition**. *arXiv preprint arXiv:1409.1556*, 2014.

[173] Joyce Xu. **An Intuitive Guide to Deep Network Architectures**, 2017.

[174] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. **Densely connected convolutional networks**. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[175] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. **Going deeper with convolutions**. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[176] François Chollet. **Xception: Deep learning with depthwise separable convolutions**. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[177] Min Lin, Qiang Chen, and Shuicheng Yan. **Network in network**. *arXiv preprint arXiv:1312.4400*, 2013.

[178] Chigozie Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. **Activation functions: Comparison of trends in practice and research for deep learning**. *arXiv preprint arXiv:1811.03378*, 2018.

[179] Kingma Da. **A method for stochastic optimization**. *arXiv preprint arXiv:1412.6980*, 2014.

[180] Eduardo Fernandez-Moral, Renato Martins, Denis Wolf, and Patrick Rives. **A new metric for evaluating semantic segmentation: leveraging global and contour accuracy**. pages 1051–1056. IEEE, 2018.

[181] Jonathan Long, Evan Shelhamer, and Trevor Darrell. **Fully convolutional networks for semantic segmentation**. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[182] Sylvain Arlot, Alain Celisse, et al. **A survey of cross-validation procedures for model selection**. *Statistics surveys*, **4**:40–79, 2010.

[183] Cullen Schaffer. **Selecting a classification method by cross-validation**. *Machine Learning*, **13**(1):135–143, 1993.

[184] Davide Chicco and Giuseppe Jurman. **The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation**. *BMC genomics*, **21**(1):1–13, 2020.

[185] Rehan Ashraf, Mudassar Ahmed, Sohail Jabbar, Shehzad Khalid, Awais Ahmad, Sadia Din, and Gwangil Jeon. **Content based image retrieval by using color descriptor and discrete wavelet transform**. *Journal of medical systems*, **42**(3):1–12, 2018.

[186] Rehan Ashraf, Mudassar Ahmed, Usman Ahmad, Muhammad Asif Habib, Sohail Jabbar, and Kashif Naseer. **MDCBIR-MF: multimedia data for content-based image retrieval by using multiple features**. *Multimedia tools and applications*, **79**(13):8553–8579, 2020.

[187] Yogita Mistry, DT Ingole, and MD Ingole. **Content based image retrieval using hybrid features and various distance metric**. *Journal of Electrical Systems and Information Technology*, **5**(3):874–888, 2018.

[188] Khawaja Tehseen Ahmed, Shahida Ummesafi, and Amjad Iqbal. **Content based image retrieval using image features information fusion**. *Information Fusion*, **51**:76–99, 2019.

[189] Peizhong Liu, Jing-Ming Guo, Kosin Chamnongthai, and Heri Prasetyo. **Fusion of color histogram and LBP-based features for texture image retrieval and classification**. *Information Sciences*, **390**:95–111, 2017.

[190] Wengang Zhou, Houqiang Li, Jian Sun, and Qi Tian. **Collaborative index embedding for image retrieval**. *IEEE transactions on pattern analysis and machine intelligence*, **40**(5):1154–1166, 2017.

[191] Chaorong Li, Yuanyuan Huang, and Lihong Zhu. **Color texture image retrieval based on Gaussian copula models of Gabor wavelets**. *Pattern Recognition*, **64**:118–129, 2017.

[192] Hee-Hyung Bu, Nam-Chul Kim, Chae-Joo Moon, and Jong-Hwa Kim. **Content-based image retrieval using combined color and texture features extracted by multi-resolution multi-direction filtering**. *Journal of information processing systems*, **13**(3):464–475, 2017.

[193] Atif Nazir, Rehan Ashraf, Talha Hamdani, and Nouman Ali. **Content based image retrieval system by using HSV color histogram, discrete wavelet transform and edge histogram descriptor**. In *2018 international conference on computing, mathematics and engineering technologies (iCoMET)*, pages 1–6. IEEE, 2018.

[194] Li-Wei Kang, Chao-Yung Hsu, Hung-Wei Chen, Chun-Shien Lu, Chih-Yang Lin, and Soo-Chang Pei. **Feature-based sparse representation for image similarity assessment**. *IEEE Transactions on multimedia*, **13**(5):1019–1030, 2011.

REFERENCES

[195] Zhong-Qiu Zhao, Hervé Glotin, Zhao Xie, Jun Gao, and Xindong Wu. **Cooperative sparse representation in two opposite directions for semi-supervised image annotation**. *IEEE Transactions on Image Processing*, **21**(9):4218–4231, 2012.

[196] Jayaraman J Thiagarajan, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, and Andreas Spanias. **Supervised local sparse coding of sub-image features for image retrieval**. In *2012 19th IEEE International Conference on Image Processing*, pages 3117–3120. IEEE, 2012.

[197] Chaoqun Hong and Jianke Zhu. **Hypergraph-based multi-example ranking with sparse representation for transductive learning image retrieval**. *Neurocomputing*, **101**:94–103, 2013.

[198] Dayong Wang, Steven CH Hoi, Ying He, Jianke Zhu, Tao Mei, and Jiebo Luo. **Retrieval-based face annotation by weak label regularized local coordinate coding**. *IEEE transactions on pattern analysis and machine intelligence*, **36**(3):550–563, 2013.

[199] M Srinivas, R Ramu Naidu, Challa S Sastry, and C Krishna Mohan. **Content based medical image retrieval using dictionary learning**. *Neurocomputing*, **168**:880–895, 2015.

[200] Sajad Mohamadzadeh and Hassan Farsi. **Content-based image retrieval system via sparse representation**. *IET Computer Vision*, **10**(1):95–102, 2016.

[201] Qiang Li, Yahong Han, and Jianwu Dang. **Sketch4Image: a novel framework for sketch-based image retrieval based on product quantization with coding residuals**. *Multimedia Tools and Applications*, **75**(5):2419–2434, 2016.

[202] Hao Wu, Rongfang Bie, Junqi Guo, Xin Meng, and Shenling Wang. **Sparse coding based few learning instances for image retrieval**. *Multimedia Tools and Applications*, **78**(5):6033–6047, 2019.

[203] Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. **Context-aware local binary feature learning for face recognition**. *IEEE transactions on pattern analysis and machine intelligence*, **40**(5):1139–1153, 2017.

[204] Chunjie Zhang, Jian Cheng, Jing Liu, Junbiao Pang, Qingming Huang, and Qi Tian. **Beyond explicit codebook generation: Visual representation using implicitly transferred codebooks**. *IEEE Transactions on Image Processing*, **24**(12):5777–5788, 2015.

[205] Chunjie Zhang, Chao Liang, Liang Li, Jing Liu, Qingming Huang, and Qi Tian. **Fine-grained image classification via low-rank sparse coding with general and class-specific codebooks**. *IEEE transactions on neural networks and learning systems*, **28**(7):1550–1559, 2016.

[206] Chunjie Zhang, Jian Cheng, and Qi Tian. **Structured weak semantic space construction for visual categorization**. *IEEE transactions on neural networks and learning systems*, **29**(8):3442–3451, 2017.

[207] Chunjie Zhang, Jian Cheng, and Qi Tian. **Semantically modeling of object and context for categorization**. *IEEE transactions on neural networks and learning systems*, **30**(4):1013–1024, 2018.

[208] Chunjie Zhang, Jian Cheng, and Qi Tian. **Unsupervised and semi-supervised image classification with weak semantic consistency**. *IEEE Transactions on Multimedia*, **21**(10):2482–2491, 2019.

[209] Xiaoshuang Shi, Manish Sapkota, Fuyong Xing, Fujun Liu, Lei Cui, and Lin Yang. **Pairwise based deep ranking hashing for histopathology image classification and retrieval**. *Pattern Recognition*, **81**:14–22, 2018.

[210] Lei Zhu, Jialie Shen, Liang Xie, and Zhiyong Cheng. **Unsupervised visual hashing with semantic assistant for content-based image retrieval**. *IEEE Transactions on Knowledge and Data Engineering*, **29**(2):472–486, 2016.

[211] Ahmad Alzu'bi, Abbes Amira, and Naeem Ramzan. **Content-based image retrieval with compact deep convolutional features**. *Neurocomputing*, **249**:95–105, 2017.

[212] Huei-Fang Yang, Kevin Lin, and Chu-Song Chen. **Supervised learning of semantics-preserving hash via deep convolutional neural networks**. *IEEE transactions on pattern analysis and machine intelligence*, **40**(2):437–451, 2017.

[213] Chunjie Zhang, Jian Cheng, and Qi Tian. **Multiview label sharing for visual representations and classifications**. *IEEE Transactions on Multimedia*, **20**(4):903–913, 2017.

[214] Chunjie Zhang, Jian Cheng, and Qi Tian. **Multiview semantic representation for visual recognition**. *IEEE transactions on cybernetics*, **50**(5):2038–2049, 2018.

[215] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. **ImageNet classification with deep convolutional neural networks**. *Communications of the ACM*, **60**(6):84–90, 2017.

[216] Yi Sun, Xiaogang Wang, and Xiaoou Tang. **Deep learning face representation from predicting 10,000 classes**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1891–1898, 2014.

[217] Andrej Karpathy and Li Fei-Fei. **Deep visual-semantic alignments for generating image descriptions**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[218] Zechao Li, Jinhui Tang, and Tao Mei. **Deep collaborative embedding for social image understanding**. *IEEE transactions on pattern analysis and machine intelligence*, **41**(9):2070–2083, 2018.

[219] Nikolaos Kondylidis, Maria Tzelepi, and Anastasios Tefas. **Exploiting tf-idf in deep convolutional neural networks for content based image retrieval**. *Multimedia Tools and Applications*, **77**(23):30729–30748, 2018.

[220] Jennifer Rowley. **The wisdom hierarchy: representations of the DIKW hierarchy**. *Journal of information science*, **33**(2):163–180, 2007.

[221] Karl R Popper and Milton Amado. *A sociedade aberta e seus inimigos*. Itatiaia, 1998.

[222] JJW Roche, Russell T Wenn, Opinder Sahota, and Christopher G Moran. **Effect of comorbidities and postoperative complications on mortality after hip fracture in elderly people: prospective observational cohort study**. *Bmj*, **331**(7529):1374, 2005.

[223] Delia Codruta Rogozan. *Gestion de l'évolution des ontologies: méthodes et outils pour un référencement sémantique évolutif fondé sur une analyse des changements entre versions d'ontologie*. PhD thesis, Télé-université, 2008.

[224] Paul Robert. **Le petit robert**. *Paris, Ed. Le Petit Robert*, 1990.

[225] Robert Neches, Richard E Fikes, Tim Finin, Thomas Gruber, Ramesh Patil, Ted Senator, and William R Swartout. **Enabling technology for knowledge sharing**. *AI magazine*, **12**(3):36–36, 1991.

[226] Pim Borst, Hans Akkermans, and Jan Top. **Engineering ontologies**. *International journal of human-computer studies*, **46**(2-3):365–406, 1997.

[227] Rudi Studer, V Richard Benjamins, and Dieter Fensel. **Knowledge engineering: Principles and methods**. *Data & knowledge engineering*, **25**(1-2):161–197, 1998.

[228] B Hassett-Sipple, J Swartout, and R Schoeny. **Mercury study report to Congress. Volume 5. Health effects of mercury and mercury compounds**. Technical report, Environmental Protection Agency, Research Triangle Park, NC (United States ..., 1997.

[229] John F Sowa. **Top-level ontological categories**. *International journal of human-computer studies*, **43**(5-6):669–685, 1995.

[230] Gertjan Van Heijst, A Th Schreiber, and Bob J Wielinga. **Using explicit ontologies in KBS development**. *International journal of human-computer studies*, **46**(2-3):183–292, 1997.

[231] Michael Uschold, Michael Gruninger, et al. **Ontologies: Principles, methods and applications**. *TECHNICAL REPORT-UNIVERSITY OF EDINBURGH ARTIFICIAL INTELLIGENCE APPLICATIONS INSTITUTE AIAI TR*, 1996.

[232] Nicola Guarino. **Understanding, building and using ontologies**. *International journal of human-computer studies*, **46**(2-3):293–310, 1997.

[233] Nicola Guarino and Christopher Welty. **A formal ontology of properties**. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 97–112. Springer, 2000.

[234] Alexander Maedche and Steffen Staab. **Discovering conceptual relations from text**. In *Ecai*, **321**, page 27, 2000.

[235] Ouafek Naouel. *Les ontologies spatiales: la génération automatique d'ontologies Spatiales à partir du modèle conceptuel spatio-temporel MADS*. PhD thesis, thèse de Magister, Université Mentouri de Constantine, 28/04, 2009.

[236] Tim Berners-Lee, James Hendler, and Ora Lassila. **The semantic web**. *Scientific american*, **284**(5):34–43, 2001.

[237] Wilson Wong, Wei Liu, and Mohammed Bennamoun. *Ontology learning and knowledge discovery using the web: challenges and recent advances*. Information Science Reference Hershey, PA, 2011.

[238] Hans-Jörg Happel and Stefan Seedorf. **Applications of ontologies in software engineering**. In *Proc. of Workshop on Sematic Web Enabled Software Engineering"(SWESE) on the ISWC*, pages 5–9. Citeseer, 2006.

[239] OMG Business Process Modeling Notation Specification. **Object management group**. *Needham, MA, USA*, **2**(2), 2006.

[240] Charles J Fillmore and Beryl T Atkins. **Toward a frame-based lexicon: The semantics of RISK and its neighbors**. *Frames, fields and contrasts: New essays in semantic and lexical organization*, **75**:102, 1992.

[241] Charles J Fillmore and Collin Baker. **A frames approach to semantic analysis**. In *The Oxford handbook of linguistic analysis*. 2010.

[242] George A Miller. **WordNet: a lexical database for English**. *Communications of the ACM*, **38**(11):39–41, 1995.

[243] Christiane Fellbaum et al. **Wordnet: An electronic lexical database mit press**. *Cambridge, Massachusetts*, 1998.

[244] NV Loukachevitch. **Thesauri in information retrieval tasks [ ]**, 2011.

[245] BV Dobrov and NV Lukashevich. **RuTez thesaurus as a resource for solving information retrieval problems**. In *Knowledge - Ontologies - Theories. - 2009 - Access mode: http: // math. nsc. ru / conference / zont09 / reports / 93Dobrov-Lukashevich. pdf (date accessed: 01.04.2013)*, 2009.

[246] Roberto Navigli and Simone Paolo Ponzetto. **BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network**. *Artificial intelligence*, **193**:217–250, 2012.

[247] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. *Ontology learning from text: methods, evaluation and applications*, **123**. IOS press, 2005.

[248] Thibault Mondary, Sylvie Després, Adeline Nazarenko, and Sylvie Szulman. **Construction d'ontologies à partir de textes: la phase de conceptualisation**. In *19èmes Journées Francophones d'Ingénierie des Connaissances (IC 2008)*, pages 87–98, 2008.

[249] Jennifer L Zamanian, Lijun Xu, Lynette C Foo, Navid Nouri, Lu Zhou, Rona G Giffard, and Ben A Barres. **Genomic analysis of reactive astrogliosis**. *Journal of neuroscience*, **32**(18):6391–6410, 2012.

[250] Cécile Fabre and Didier Bourigault. **Linguistic clues for corpus-based acquisition of lexical dependencies**. *Proceeding of Corpus Linguistics, Lancaster*, 2001.

[251] Marie-Claude L'homme. *La terminologie: principes et techniques*. Pum, 2004.

[252] Erik Neveu. *Les voyages des cultural studies*. Number 187-188. Editions de l'EHESS, 2008.

[253] Chris Biemann. **Ontology learning from text: A survey of methods.** In *LDV forum*, **20**, pages 75–93, 2005.

[254] Olena Orobinska, Jean-Hugues Chauchat, and Natalya Sharonova. **Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security)**. In *Computational linguistics andintelligent systems (COLINS 2017)*. National Technical University «KhPI», 2017.

[255] Haïm Brezis and Xavier Cabré. **Some simple nonlinear PDE's without solutions**. *Bollettino della Unione Matematica Italiana-B*, **1**(2):223–262, 1998.

[256] Éric Gaussier and François Yvon. *Modèles statistiques pour l'accès à l'information textuelle*. Lavoisier, 2011.

[257] Kenneth Ward Church and William A Gale. **Poisson mixtures.** *Nat. Lang. Eng.*, **1**(2):163–190, 1995.

[258] Francesco Sclano and Paola Velardi. **Termextractor: a web application to learn the shared terminology of emergent web communities**. In *Enterprise Interoperability II*, pages 287–290. Springer, 2007.

[259] Inderjit S Dhillon, Subramanyam Mallela, and Dharmendra S Modha. **Information-theoretic co-clustering**. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 89–98, 2003.

[260] Khurshid Ahmad, Lee Gillam, Lena Tostevin, et al. **University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder)**. In *TREC*, pages 1–8, 1999.

[261] Anselmo Peñas, Felisa Verdejo, Julio Gonzalo, et al. **Corpus-based terminology extraction applied to information access**. In *Proceedings of Corpus Linguistics*, **2001**, page 458, 2001.

REFERENCES

[262] Daniela Kurz and Feiyu Xu. **Text mining for the extraction of domain relevant terms and term collocations**. In *Proceedings of the International Workshop on Computational Approaches to Collocations*. Citeseer, 2002.

[263] Roberto Basili, Alessandro Moschitti, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. **A contrastive approach to term extraction**. In *TIA 2001: terminologie et intelligence artificielle (Nancy, 3-4 mai 2001)*, pages 119–128, 2001.

[264] Wilson Wong, Wei Liu, and Mohammed Bennamoun. **Determining termhood for learning domain ontologies using domain prevalence and tendency**. In *Proceedings of the sixth Australasian conference on Data mining and analytics-Volume 70*, pages 47–54. Citeseer, 2007.

[265] Alexander Gelbukh, Grigori Sidorov, Eduardo Lavin-Villa, and Liliana Chanona-Hernandez. **Automatic term extraction using log-likelihood based comparison with general reference corpus**. In *International conference on application of natural language to information systems*, pages 248–255. Springer, 2010.

[266] S Kullback. **Information Theory and Statistics–Dover Edition**, 1997.

[267] Wen Zhang, Taketoshi Yoshida, Xijin Tang, and Tu-Bao Ho. **Improving effectiveness of mutual information for substantival multiword expression extraction**. *Expert Systems with Applications*, **36**(8):10919–10930, 2009.

[268] Gerlof Bouma. **Normalized (pointwise) mutual information in collocation extraction**. *Proceedings of GSCL*, pages 31–40, 2009.

[269] Béatrice Daille. *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. PhD thesis, Paris 7, 1994.

[270] Joaquim Ferreira Da Silva, Gabriel Pereira Lopes, Quinta Da Torre, and Monte Da Caparica. **A local maxima method and a fair dispersion normalization for extracting multiword units**. In *In Proceedings of the 6th Meeting on the Mathematics of Language*. Citeseer, 1999.

[271] Kathleen McKeown, Frank Smadja, and Vasileios Hatzivassiloglou. **Translating collocations for bilingual lexicons: A statistical approach**. 1996.

[272] Mihoko Kitamura and Yuji Matsumoto. **Automatic extraction of word sequence correspondences in parallel corpora**. In *Fourth Workshop on Very Large Corpora*, 1996.

[273] Youngja Park, Roy J Byrd, and Branimir Boguraev. **Automatic Glossary Extraction: Beyond Terminology Identification.** In *COLING*, **10**, pages 1072228–1072370, 2002.

[274] Vidas Daudaravičius and Rūta Marcinkevičienė. **Gravity counts for the boundaries of collocations**. *International Journal of Corpus Linguistics*, **9**(2):321–348, 2004.

[275] Ted E Dunning. **Accurate methods for the statistics of surprise and coincidence**. *Computational linguistics*, **19**(1):61–74, 1993.

[276] Sophia Ananiadou. **A methodology for automatic term recognition**. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*, 1994.

[277] Katerina T Frantzi and Sophia Ananiadou. **Automatic term recognition using contextual cues**. In *In Proceedings of 3rd DELOS Workshop*. Citeseer, 1997.

[278] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. **Automatic recognition of multi-word terms:. the c-value/nc-value method**. *International journal on digital libraries*, **3**(2):115–130, 2000.

[279] Michael Nokel, EI Bolshakova, and Natalia Loukachevitch. **Combining multiple features for single-word term extraction**. *Proceedings of Dialog*, pages 490–501, 2012.

[280] Hirosi Nakagawa and Tatsunori Mori. **Automatic term recognition based on statistics of compound nouns and their components**. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, **9**(2):201–219, 2003.

[281] Michael Lesk. **Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone**. *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, 1986.

[282] Paul Buitelaar and Bogdan Sacaleanu. **Extending synsets with medical terms**. *Proceedings of the First International WordNet Conference*, **324**, 2002.

[283] Joerg-Uwe Kietz, Alexander Maedche, and Raphael Volz. **A method for semi-automatic ontology acquisition from a corporate intranet**. *EKAW-2000 Workshop "Ontologies and Text", Juan-Les-Pins, France, October 2000*, 2000.

[284] Bernardo Magnini and Carlo Strapparava. **Experiments in word domain disambiguation for parallel texts**. *ACL-2000 Workshop on Word Senses and Multi-linguality*, pages 27–33, 2000.

[285] Roberto Navigli and Paola Velardi. **Learning domain ontologies from document warehouses and dedicated web sites**. *Computational Linguistics*, **30**(2):151–179, 2004.

[286] Zellig Sabbettai Harris. **Mathematical structures of language**. 1968.

[287] Thomas K Landauer and Susan T Dumais. **A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.** *Psychological review*, **104**(2):211, 1997.

[288] Marco Baroni and Sabrina Bisi. **Using cooccurrence statistics and the Web to discover synonyms in a technical language.** *LREC*, 2004.

[289] Peter D Turney. **Mining the web for synonyms: PMI-IR versus LSA on TOEFL**. *European conference on machine learning*, pages 491–502, 2001.

[290] Richard Evans and Stafford Street. **A framework for named entity recognition in the open domain**. *Recent advances in natural language processing III: selected papers from RANLP*, **260**(267-274):110, 2003.

[291] Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. **Web-scale information extraction in knowitall: (preliminary results)**. *Proceedings of the 13th international conference on World Wide Web*, pages 100–110, 2004.

[292] Marti A Hearst. **Automatic acquisition of hyponyms from large text corpora**. *Coling 1992 volume 2: The 15th international conference on computational linguistics*, 1992.

[293] Paul Buitelaar, Daniel Olejnik, and Michael Sintek. **A protégé plug-in for ontology extraction from text based on linguistic analysis**. *European Semantic Web Symposium*, pages 31–44, 2004.

[294] David Faure and Claire Nédellec. **A corpus-based conceptual clustering method for verb frames and ontology acquisition**. *LREC workshop on adapting lexical and corpus resources to sublanguages and applications*, **707**(728):30, 1998.

[295] Mark Sanderson and Bruce Croft. **Deriving concept hierarchies from text**. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213, 1999.

[296] Bruno Bachimont. **Engagement sémantique et engagement ontologique: conception et réalisation d'ontologies en ingénierie des connaissances**. *Ingénierie des connaissances: évolutions récentes et nouveaux défis*, pages 305–323, 2000.

[297] Philipp Cimiano, Andreas Hotho, and Steffen Staab. **Learning concept hierarchies from text corpora using formal concept analysis**. *Journal of artificial intelligence research*, **24**:305–339, 2005.

[298] Bryan G Fry, Nicholas R Casewell, Wolfgang Wüster, Nicolas Vidal, Bruce Young, and Timothy NW Jackson. **The structural and functional diversification of the Toxicofera reptile venom system**. *Toxicon*, **60**(4):434–448, 2012.

[299] Ido Dagan, Oren Glickman, and Bernardo Magnini. **The pascal recognising textual entailment challenge**. *Machine Learning Challenges Workshop*, pages 177–190, 2005.

[300] ADAM FARQUHAR, RICHARD FIKES, and JAMES RICE. **The Ontolingua Server: a tool for collaborative ontology construction**. *International Journal of Human-Computer Studies*, **46**(6):707–727, 1997.

[301] York Sure, Juergen Angele, and Steffen Staab. **OntoEdit: Multifaceted inferencing for ontology engineering**. pages 128–152. Springer, 2003.

[302] John H Gennari, Mark A Musen, Ray W Fergerson, William E Grosso, Monica Crubézy, Henrik Eriksson, Natalya F Noy, and Samson W Tu. **The evolution of Protégé: an environment for knowledge-based systems development**. *International Journal of Human-computer studies*, **58**(1):89–123, 2003.

[303] Natalya Fridman Noy, Monica Crubézy, Ray W Fergerson, Holger Knublauch, Samson W Tu, Jennifer Vendetti, and Mark A Musen. **Protégé-2000: an open-source ontology-development and knowledge-acquisition environment.** *AMIA... Annual Symposium proceedings. AMIA Symposium*, **2003**:953–953, 2003.

[304] Peter Haase, Holger Lewen, Rudi Studer, Duc Thanh Tran, Michael Erdmann, Mathieu d'Aquin, and Enrico Motta. **The neon ontology engineering toolkit**. *WWW*, 2008.

[305] Leslie F Sikos. **Description logics in multimedia reasoning**. 2017.

[306] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. **The open images dataset v4**. *International Journal of Computer Vision*, pages 1–26, 2020.

[307] Kashif Ahmad, Nicola Conci, Giulia Boato, and Francesco GB De Natale. **USED: a large-scale social event detection dataset**. pages 1–6, 2016.

[308] Unaiza Ahsan, Chen Sun, and Irfan Essa. **DiscrimNet: Semi-Supervised Action Recognition from Videos using Generative Adversarial Networks**. *arXiv preprint arXiv:1801.07230*, 2018.

[309] L. Li and Li Fei-Fei. **What, where and who? Classifying events by scene and object recognition**. pages 1–8, 2007.

[310] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzàlez, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon. **ChaLearn Looking at People 2015: Apparent Age and Cultural Event Recognition Datasets and Results**. *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 243–251, 2015.

[311] Yuanjun Xiong, Kai Zhu, Dahua Lin, and X. Tang. **Recognize complex events from static images by fusing deep channels**. pages 1600–1609, 2015.

[312] Fehima Achour, Emna Bouazizi, and Wassim Jaziri. **Improving the quality of service of real-time database systems through a semantics-based scheduling strategy**. *International Journal of Intelligent Information and Database Systems*, **14**(1):96–114, 2021.

[313] Vijay Shrinath Patil and Pramod Jagan Deore. **Semantic image retrieval using random forest-based AdaBoost learning**. *International Journal of Intelligent Information and Database Systems*, **12**(3):229–243, 2019.

[314] Dwijen Rudrapal and Amitava Das. **Semantic role labelling of English tweets through sentence boundary detection**. *International Journal of Intelligent Information and Database Systems*, **11**(4):225–235, 2018.

[315] P Gayathri and Natarajan Jaisankar. **A hybrid neuro-fuzzy system-based ranking function and its application to effective medical information retrieval**. *International Journal of Intelligent Information and Database Systems*, **9**(3-4):248–268, 2016.

[316] Khalid Anwar, Jamshed Siddiqui, and Shahab Saquib Sohail. **Machine learning-based book recommender system: a survey and new perspectives**. *International Journal of Intelligent Information and Database Systems*, **13**(2-4):231–248, 2020.

[317] Aditya Khamparia, Sanjay Kumar Singh, Ashish Kr Luhach, and Xiao-Zhi Gao. **Classification and analysis of users review using different classification techniques in intelligent e-learning system**. *International Journal of Intelligent Information and Database Systems*, **13**(2-4):139–149, 2020.

[318] Huaiguang Wu, Daiyi Li, and Ming Cheng. **Chinese text classification based on character-level CNN and SVM**. *International Journal of Intelligent Information and Database Systems*, **12**(3):212–228, 2019.

[319] Liang Chen, Shuo Xu, Lijun Zhu, Jing Zhang, Xiaoping Lei, and Guancan Yang. **A deep learning based method for extracting semantic information from patent documents**. *Scientometrics*, **125**(1):289–312, 2020.

[320] Jalila Filali, Hajer Baazaoui-Zghal, and Jean Martinet. **OntoAnnClass: Ontology-Based Image Annotation driven by Classification using HMAX features**. *Multimedia Tools and Applications*, 2020.

[321] Eric Muller-Budack, Matthias Springstein, Sherzod Hakimov, Kevin Mrutzek, and Ralph Ewerth. **Ontology-driven event type classification in images**. pages 2928–2938, 2021.

[322] Lorenzo Torresani, Martin Szummer, and Andrew W. Fitzgibbon. **Efficient Object Category Recognition Using Classemes**. *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I*, **6311**:776–789, 2010.

[323] Muhammad Nabeel Asim, Muhammad Wasim, Muhammad Usman Ghani Khan, Waqar Mahmood, and Hafiza Mahnoor Abbasi. **A survey of ontology learning techniques and applications**. *Database*, **2018**, 2018.

[324] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[325] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. **Faster r-cnn: Towards real-time object detection with region proposal networks**. *IEEE transactions on pattern analysis and machine intelligence*, **39**(6):1137–1149, 2016.

[326] Mouna Ben Ishak, Philippe Leray, and Nahla Ben Amor. **Ontology-based generation of object oriented bayesian networks**. 2011.

[327] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. **Factorizable net: an efficient subgraph-based framework for scene graph generation**. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–351, 2018.

[328] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. **R-FCN: Object Detection via Region-Based Fully Convolutional Networks**. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 379–387, 2016.

[329] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. **Deformable Convolutional Networks**. Oct 2017.

[330] Tongxin Hu, Wentong Liao, Michael Ying Yang, and Bodo Rosenhahn. **Exploiting Attention for Visual Relationship Detection**. *German Conference on Pattern Recognition*, pages 331–344, 2019.

[331] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. **Visual relationship detection with language priors**. In *European conference on computer vision*, pages 852–869. Springer, 2016.

[332] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. **Visual genome: Connecting language and vision using crowdsourced dense image annotations**. *International journal of computer vision*, **123**(1):32–73, 2017.

[333] Marek Kubis. **A query language for WordNet-like lexical databases**. *International Journal of Intelligent Information and Database Systems*, **9**(2):103–133, 2016.

[334] Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. **Understanding, categorizing and predicting semantic image-text relations**. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 168–176, 2019.

[335] Qingyong Li, Siwei Luo, and Zhongzhi Shi. **Fuzzy aesthetic semantics description and extraction for art image retrieval**. *Computers & Mathematics with Applications*, **57**(6):1000–1009, 2009.

[336] Christian Henning and Ralph Ewerth. **Estimating the information gap between textual and visual representations**. *International journal of multimedia information retrieval*, **7**(1):43–56, 2018.