Multimedia Tools and Applications (2020) 79:21679–21741
https://doi.org/10.1007/s11042-020-08862-1

A review on visual content-based and users' tags-based image annotation: methods and techniques

Mariam Bouchakwa[1] · Yassine Ayadi[1] · Ikram Amous[1]

Received: 3 January 2019 / Revised: 20 February 2020 / Accepted: 18 March 2020 /
Published online: 9 May 2020

Abstract
In the current era of digital communication, the use of images is growing exponentially since they are one of the best ways of expressing, sharing and memorizing knowledge. In fact, images can be used in various real-world applications, like biology, medical diagnosis, space research, remote sensing, etc. However, finding the most relevant images that meet the users' needs is a challenging task, especially when the search is performed over gigantic amounts of images. This has led to the emergence of several image retrieval studies during the past two decades. Typically, research studies in this area were focused on the *Content-based Image Retrieval* (CBIR). However, extensive research have proved that there is a '*semantic gap*' between the visual information captured by the imaging devices and the image semantics understandable by humans. As an alternative, researchers' efforts have been oriented towards the *Text-based Image Retrieval* (TBIR). Indeed, TBIR is a typical method that helps bridge the issue of '*semantic gap*' between the low-level image features and the high-level image semantics. Its policy consists in associating textual descriptions with the images, which constitute the focus of the research queries later on. In this paper, we analyze various image annotation methods, namely: *Visual Content-based* and *Users' Tags-based Image Annotation Methods*. In particular, we focus on the visual content-based image annotation techniques since they are one of the dynamic research fields nowadays.

Keywords Image annotation · Image segmentation · Feature extraction · Machine learning · Deep learning · Semantic measure

✉ Mariam Bouchakwa
mariam.bouchekwa@gmail.com

Yassine Ayadi
ayadi.yassine@gmail.com

Ikram Amous
Ikram.amous@enetcom.usf.tn

[1] MIRACL Laboratory, Technopole of Sfax, University of Sfax, P.O.Box 242, 3031 Sfax, Tunisia

1 Introduction

In the early days of digital communication, the exchange of information was mainly used through text documents. Nowadays, and especially with the progress of multimedia technologies and the emergence of Web 2.0, vast amounts of information are stored in a visual form. This has created an urgent need for effective and efficient tools that help find the required visual information. Therefore, a large number of research studies that focus on image retrieval have emerged over the past two decades. These research efforts can be divided into two types of approaches:

1.1 The content-based image retrieval (CBIR)

This type of approach focuses on the retrieval of images based on their visual characteristics, such as the shape, texture and color [149]. It involves three steps as shown in Fig. 1:

Offline feature extraction step It consists in representing the set of images of a collection according to their visual contents. Thus, each image is indexed in a condensed form represented by visual characteristic vectors, color and texture histograms, etc.

Online feature extraction step It consists in extracting the visual content of the query image and representing it by visual characteristic vectors, color and texture histograms, etc.

Similarity matching step It consists in matching the visual features of the query image with those of the collection images in order to achieve the results that are visually similar.

1.2 Text-based image retrieval (TBIR)

This type of approach focuses on the return of potentially relevant images compared to a user-described textual research query [335]. It involves three steps as shown in Fig. 2:

Image annotation step It consists in associating one or more keywords to each image stored in the database.

Search formulation step It consists in defining a set of formulas in order to reformulate the query initially sent by a user.

Matching step It consists in matching the image keywords with the reformulated query in order to achieve the results that meet the user's needs.

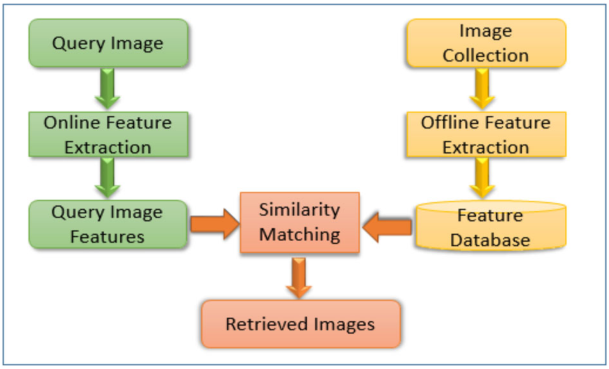

Fig. 1 General architecture of CBIR systems

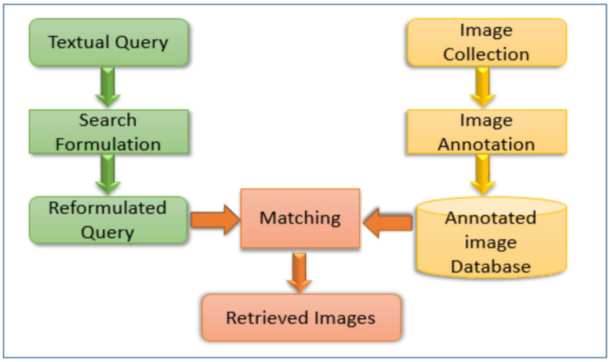

Fig. 2 General architecture of **TBIR** systems

In the last decades, the CBIR technology has been used in general-purpose applications [15, 34, 39, 41, 103, 114, 115, 162, 195, 206, 305, 329] and specific domains, such as medical sciences [23, 209], robotics [141] and remote sensing [82]. Thus, it has been the focus of numerous surveys in the literature [61, 160, 169, 173, 242, 257, 287, 288]. Indeed, the CBIR technology is frequently adopted because of its ability to provide more than one related outcomes occur by only one search if more than one equally likely image present in the database. The required time also is less to find all these related images. In addition, the feature extraction methods are easy, effective and less expensive. However, this technology suffers from two major limitations: First, it is impractical for users to use CBIR systems because they need to provide query images. Second, the '*semantic gap*' between the low-level content features and the semantic concepts used by humans to interpret images make the use of CBIR systems a challenging task [261]. As defined in [261], the '*semantic gap*' problem is '*the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation*'. Thus, semantically different images may have similar low-level characteristics. The example of the two images in Fig. 3 illustrates the '*semantic gap'* issue. Indeed, these images have similar color histograms, but semantically they are different: in the foreground of the first image include flowers, while the second image shows a man playing golf [218].

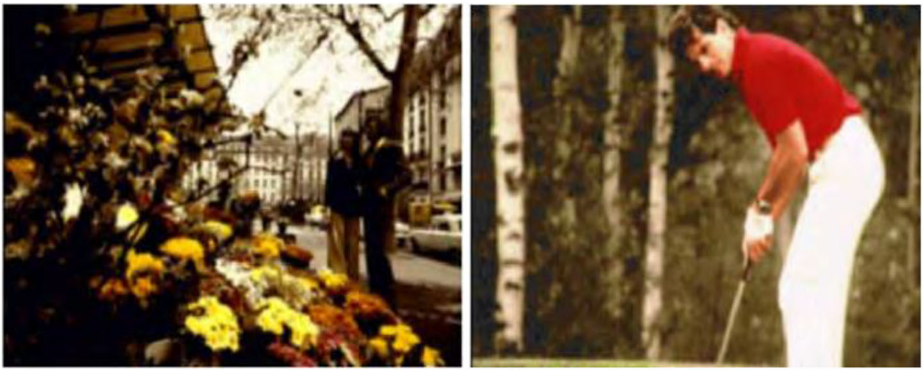

Fig. 3 Two images with similar colors histograms, but two different semantic meanings

To avoid the issues caused by the CBIR technology, most users prefer to use textual queries [261]. Therefore, the images should be annotated. In fact, the annotation process consists in associating textual descriptors (*keywords*, *labels*, *terms or captioning descriptions*) with the image sets. These descriptors will be used at the indexing phase, and will be the focus of research queries later on. In fact, the TBIR technology could lead to a lot of critical applications.

Medicine: The importance of digital image retrieval techniques increases in the emerging fields of medical imaging, picture archiving and communication systems (PACS). Up to now, textual index entries are mandatory to find medical images from hospital archives or other sources [4, 314].

Radiology: Automatic image search methods help radiologists interpret a given image by identifying similar images in databases. Recent work has shown the interest of characterizing the content of images by semantic terms [152]. These terms can be used to describe a large number of information relating to the visual content of images. They can be derived from radiologist observations or automatically predicted from low-level features extracted from pixels [151].

Digital libraries: Some general digital libraries, such as the Internet Public Library and the National Science Digital Library (NSDL), are widely known and used. In fact, the advance of computer technology makes it possible to include a colossal amount of information in different formats within a digital library. In addition to traditional text-based documents, such as books and articles, other categories of materials, including images [106], audio, and video, can also be easily digitized and stored. Therefore, an effective search of this multimedia information based on textual queries through the interface of digital libraries becomes a significant research topic.

Touristic attractions: Over the years, the number of images taken by tourists to memorize their vacation memories, unfolded events and visited places has grown exponentially thanks to the popularity of digital cameras and the integration of digital sensors into mobile phones. Thus, there is an urgent need for effective and efficient annotation policies that help retrieve as quickly as possible the relevant visual information from great collections of tourist-type images [27, 28, 196].

Image research engines and social medias: In the era of digital communication, massive amounts of information are daily exchanged between people via the *World Wide Web*. In the beginning, the exchange of information was mainly used through text documents. However, with the technological progress, the exchange of information covers different other forms, such as audio, video and image. Thanks to the advancements of digital image acquisition devices, image capturing is no longer a difficult task. In fact, images have been increasingly used since they are one of the best ways of expressing, sharing and memorizing knowledge. The retrieving of the relevant images from gigantic image collections by using textual queries attracts more and more users in various professional and amateur fields, specially through the Web engines (such as *Google* and *Yahoo*) and the social media (such as *Flickr* and *Instagram*).

As previously mentioned, the TBIR technology requires the annotation of the image sets. However, the annotation content as well as the techniques adopted to generate the annotation vocabulary constitute challenges for researchers. In fact, some research studies have focused on the annotation of social images based on the refinement of the tagging information associated with them by the social community [83, 147, 199, 231, 316, 317, 342]. Others have processed to the extraction of textual information from the visual content of images [16, 19, 65, 106, 154, 167, 181, 200, 224, 227, 244, 274, 320, 343]. Thus, recent surveys have focused in studying TBIR technology in the literature [2, 47, 210, 336]. However, none of them gives attention to the techniques used to refine the social information to annotate images. Besides, they do not provide complete studies over the automatic image annotation (AIA) techniques. In order to supplement existing reviews in the literature, we analyze in this paper various image annotation methods, namely: *Visual Content-based Image Annotation* and *Users' Tags-based Image Annotation*. We specifically studying the social tag refinement techniques and the visual content-based image annotation techniques, including image segmentation, visual feature extraction and machine/deep learning.

The remaining of this paper is organized as follows: In Sect. 2, we focus on identifying the parameters of the image annotation systems. In Sect. 3, we provide an overview on the image annotation methods. In Sect. 4, the visual content-based image annotation techniques are descripted. In Sect. 5, we discuss some challenges, open issues, and promising directions in image annotation field.

2 Parameters of image annotation systems

Image annotation systems are characterized by a set of parameters. As shown in Fig. 4, the values that can have these parameters make it possible to describe this type of systems.

The parameters of the annotation systems can be reviewed in the form of five questions:

2.1 What information should be analysed to annotate an image ?

It is about defining the set of features from which images can be analyzed and interpreted in order to generate the necessary annotations [95]. We can mention:

The visual information [8, 19, 24, 65, 119, 154] It helps annotate images by using their visual features (such as texture, color, shape, etc.).

The textual information [27, 49, 113, 120, 125] It helps annotate images by using their textual features (such as URLs, title, users' comments and tags, etc.).

These pieces of information cannot be employed in their raw form. Indeed, a treatment process is a fundamental step to generate significant annotations for images. Different image processing methods and techniques are detailed in Sect. 3 and Sect. 4, respectively.

2.2 What views should be considered to annotate an image ?

It is about defining the angles from which images can be seen and analyzed [26]. This can be realized according to several views, such as:

The structural (anatomical) view [27, 224, 244, 265, 292, 297, 298, 343] It helps define the sensory objects depicted on an image (people, buildings, monuments, means of transport, etc.) as well as their components.

The behavior view [27, 197] It helps define the activities performed within an image (sports, adventure, trade, etc.).

The event view [27, 274] It helps define the events unfolded within an image (festivals, parties, etc.).

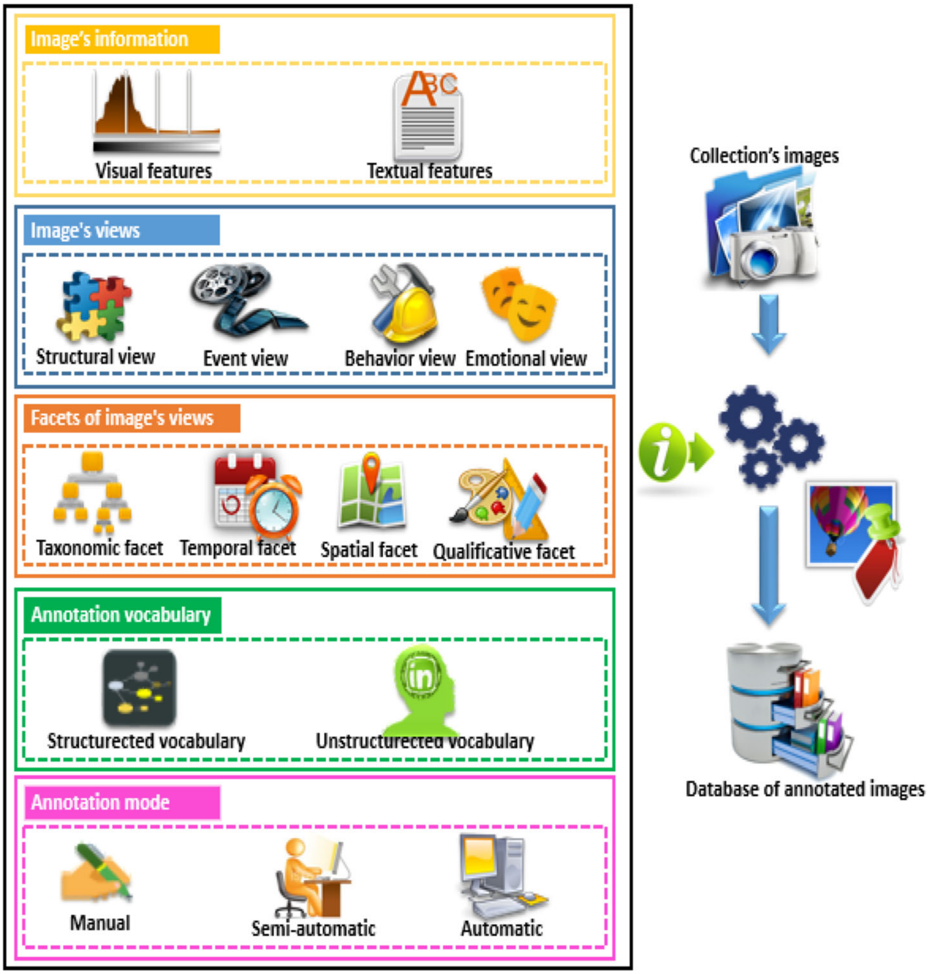

Fig. 4 Parameters of image annotation systems

The emotional view [107, 310] It helps define the emotions of the people visualized on an image (sadness, joy, fear, etc.).

2.3 What views' facets should to be descripted to annotate an image ?

It is about defining the different facets of each view of an image to be descript [26], among which we can mention:

The Taxonomic facet [19, 27, 65, 167] It helps categorize the physical objects depicted on the images, the unfolding activities and events, the characters' emotions, etc. For instance, the taxonomic facet of the visual object '*Airplane*' is '*Air Transportation Service*' and that of the visual object '*Bus*' is '*Ground Transportation Service*'.

The Qualificative facet [27, 167, 227, 292] It helps describe the observable specifications of the physical objects as well as those of their components, the unfolding activities and events, and distinguish them from other objects, activities and events. For example, the qualification facet of the event '*Battle of the Oranges*' is '*Fruit Festival*' and that of the event '*Las Fallas*' is '*Folkloric Festival*'.

The Spatial facet [16, 106, 200, 266] It helps indicate the location of the physical objects, and the unwinding location of the events and activities. For instance, the spatial facet of the landmark '*Sagrada Familia*' is '*Spanish Palace*' and that of the landmark '*Eiffel Tower*' is '*French Tower*'.

The Temporal facet [27, 266] It helps indicate the epoch of construction of the historical monuments, and the unwinding season of the events and activities. For example, the temporal facet of the activity '*Snowboarding*' is '*Winter skiing*' and that of the activity '*Kite surfing*' is '*Summer skiing*'.

2.4 What vocabulary should be used to annotate an image ?

It is about defining the basic source from which the vocabulary of the annotation can be selected [26]. Indeed, image annotation can rely on an *unstructured vocabulary* [19, 65, 106, 112, 113, 120, 125, 147, 170, 199, 231, 297, 298, 316, 317, 342], such as the keywords chosen and introduced by users, or on a *structured vocabulary* [4, 92, 151, 167, 181, 200, 224, 227, 244, 265, 274, 292, 343], such as the ontological terminologies. Indeed, an ontology is a controlled and structured vocabulary of agreed-upon labels (or terms) that represent the knowledge of the different entities of a particular domain [51]. The use of ontologies has gained increasing importance since the complexity, number, and size of specific field datasets have increased [12].

2.5 How to annotate an image ?

It is about defining the manner according to which the annotation process can be executed [26]. In fact, there are three available modes of executing an image annotation process:

The manual annotation [4, 92, 151, 167, 227, 244, 265, 274, 297, 298] It requires that the human annotators introduce some descriptive keywords when they browse a collection of images.

The automatic annotation [14, 27, 48, 65, 83, 86, 112, 113, 120, 125, 147, 170, 199, 202, 231, 292, 303, 306, 316] It helps automatically detect and classify the objects depicted on any image and label them with a set of keywords.

Table 1 Contrast of annotation modes

Annotation mode	Initial human interaction	Machine task	Human effort
Manual	Type some descriptive keywords.	Provide a storage space, such as a disk space or database, to save the image-associated annotations.	Provide sufficient descriptive information for retrieval purposes.
Automatic	No interaction.	Automatically generate descriptive keywords by using recognition technologies.	Verify and refine the quality of the machine final output for annotation accuracy.
Semi-automatic	Provide initial descriptions in the beginning.	Analyze human descriptions and refine them.	Provide some annotations and work with machine input.

The semi-automatic annotation [16, 106, 181, 200, 224, 343] It requires the intervention of human annotators in order to generate initial descriptions of the images. These descriptions are later refined by an annotator system in order to provide definitive descriptions of these images.

A comparison of the different annotation modes is demonstrated in Table 1.

3 Image annotation methods

Digital images are widely adopted in many fields and for multiple purposes since they are good mediums of expression, memorization and communication of information. Therefore, it is often necessary to analyze, understand and describe the semantic content of images in order to annotate them. Indeed, image annotation is not only a key step that helps optimize the quality of the search results [19, 27, 65, 106, 120, 147, 181] but also an efficiency factor of other type of applications, such as computer vision training algorithms [298], supervised machine learning [227, 265], comparative analyses [167, 297], etc. Image annotation can be based on the information derived from the image visual content processing or image-associated text processing.

3.1 Visual content-based image annotation methods

With the evolution of computer vision technologies, image processing has become a promising solution for multiple applications, such as image annotation. The keywords generated after executing an annotation process often constitute the focus of the retrieval process. These keywords can reflect the visual objects depicted on the images or can be relative to their semantic content, such as activities, events and emotions.

3.1.1 Low-level feature-based image annotation

The annotation based on the low-level features of images, like color, texture and shape, relies essentially in recognizing the objects. Indeed, an object recognition process may often rely on one of the region-based segmentation techniques. Thereafter, an identifier will be assigned to each recognized object, either automatically [8, 109, 119, 219, 292] or through the intervention of human operators [4, 92, 151, 224, 244, 265, 297, 298, 343]. Automatic image annotation (AIA) is often based on one of machine learning techniques [53, 78, 121, 216, 228, 243, 272, 278, 315, 320]. Fig. 5 illustrates the progress of the automatic low-level features-based image annotation process.

For example, the collaborative Web framework ‘*WebMedSA*’, designed by Vega et al. [292], is founded on a client-server architecture that aims at managing a big set of biomedical images. On the client-side, the role of the user is simply to send a query image. The server applies a polygonal segmentation to the received image and annotate it according to its objects’ sub-anatomies. The annotation terms are extracted from one of the used ontologies. When the server accomplishes its tasks, it sends a notice to the user that receives the image described in terms of its sub-anatomies localization (x, y, z) related to their names.

Yang et al. [320] considered the image annotation as an image classification issue, in which each keyword is treated as a distinct class label. At first, each image is segmented into up to 10 regions, where each region is represented with a low-level feature vector. The classification problem is next addressed by using a *Bayesian* framework. To preserve the in-variation of the

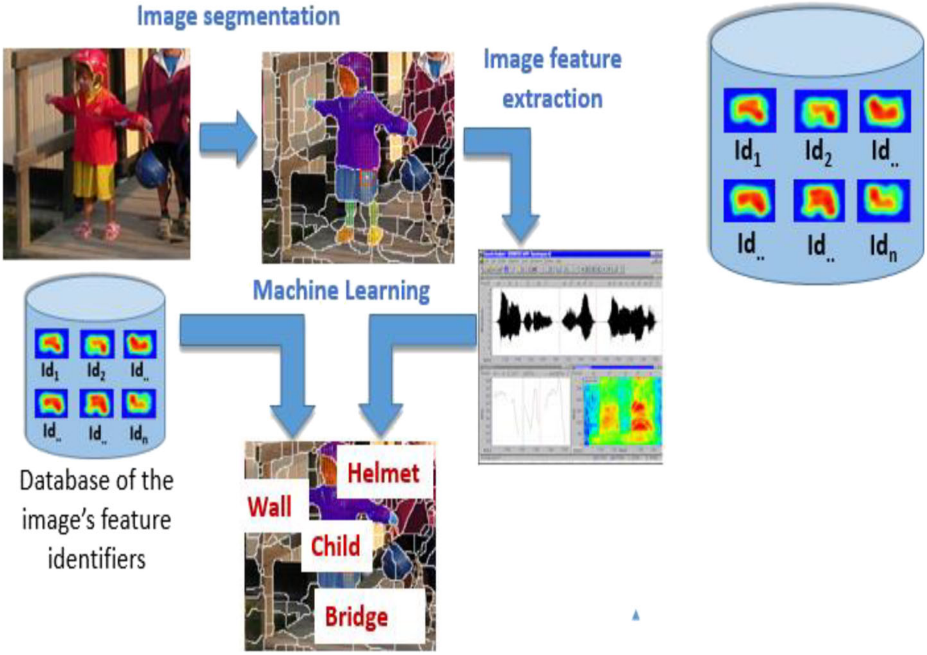

Fig. 5 Progress of the automatic low-level feature-based image annotation process

training data and reduce the noises, an estimation of the class conditional probabilities in the feature subspace is constructed via a *Complement Components Analysis* (CCA).

'*LabelMe*' is a Web-based image annotation tool designed by Russell et al. [244]. It aims at building a large database of annotated images by collecting the contributions of many people. Indeed, the users are supposed to contour and annotate the objects depicted on the images by using a set of terms derived from *WordNet*. One important concern is the quality control. This control is provided by the users themselves. The obtained image collection, with ground truth labels, is dedicated to the object detection and recognition research applications.

The '*M-OntoMat-Annotizer*' tool, suggested by Petridis et al. [224], is designed to annotate large collections of images. It is used by the information systems of research and organization, as well as the knowledge-assisted analysis of multimedia. This desktop tool has the capacity to automatically ensure the segmentation task, according to the low-level MPEG-7 visual descriptions, in order to provide detailed annotations for the objects. It covers two main functions: On the one hand, it helps users link the detected visual descriptors to the corresponding ontological terms. On the other hand, it helps enrich the domain ontologies with these multimedia descriptors, as an RDF instance form, in order to ensure an automatic reasoning afterwards.

A collaborative annotation system called '*EMERGSEM*' is suggested by Zomahoun et al. [343]. The purpose of this system is to extract the image meanings from different interpretations suggested by human annotators thanks to a domain ontology. The annotation process unfolds as follows: first, an ontology model and a lexical dictionary are proposed to the annotators. Thus, the annotators propose a set of instances, indicating the objects depicted on an image, by using the ontological concepts. Once the instances are attributed and stored as an xml file, the initial image meanings are obtained. Thereafter, a computing of meaning similarities are executed to refine the users' initial annotations. Finally, the resulting annotations are displayed and will be the focus at the research process.

A novel interactive multiscale tagging framework is proposed by Tang et al. [276]. The policy of this approach consists first in segmenting each image into multiple regions. Second, a dynamic multiscale cluster labeling strategy is proposed in order to manually label these regions, which are mapped into different buckets by efficient locality sensitive hashing (LSH) method. This step can be regarded as a coarse clustering where each bucket is a cluster. In fact, the coarse clustering can keep the efficiency of the proposed approach in dealing with large dataset. Thereafter, each cluster is recursively clustered into smaller clusters until it is able to be manually labeled with a one tag. During the labeling process, the partially obtained tags are fed back in order to refine the hashing after finish labeling several buckets. After finishing the labeling process, the region tags are combined into image tags. A tag refinement process is then performed based on a matrix decomposition method. This process is able to refine the image tags to boost their accuracy and assign tags to some unlabeled images.

3.1.2 Top-level feature-based image annotation

The annotation based on the top-level features of images aims at associating the images with a set of keywords that reflect their semantic meanings. Indeed, the semantic meaning can be related not only to the objects depicted on the images but also to their characteristics, such as localizations, taxonomies, etc. The top-level feature-based image annotation methods can be essentially divided into two categories: semantic annotation methods according to the objects depicted on the images [65, 167, 181, 227, 274] and semantic annotation methods adding spatial information [16, 19, 106, 200].

In fact, the first type of semantic annotation methods consists in detecting the visual features of images, such as the regions or objects, and deducing from them the set of global semantic meanings, such as the scenes (people, landscape, indoor, outdoor, animal, etc.) and

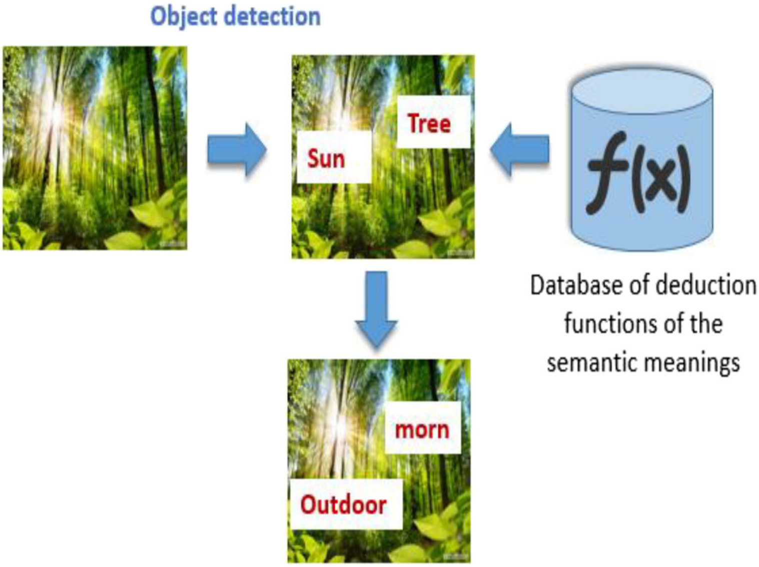

Fig. 6 Progress of the semantic annotation process using the objects depicted on an image

the taxonomies of the objects (sheep/herbivore animal, car/means of transport, etc.). Fig. 6 illustrates the progress of the semantic annotation process according to the objects depicted on an image.

For example, Magesh et al. [181] used ICONCLASS tool to extract the low-level visual features of images. Thereafter, the images are integrated within the '*protégé*' tool for the users to associate the corresponding names with the extracted objects by using the instances of the adopted ontology. Finally, the super-classes of each chosen instance as well as their semantic properties are automatically mapped and associated with the target image to enrich its initial-suggested semantic annotation.

A multi-level natural image annotation framework is suggested by Fan et al. [65] for a semantic-based retrieval. The main idea of this framework is to automatize the salient object detection [207] by using learning-oriented techniques. Indeed, the use of the salient objects permits a precise extraction of the images' features, and consequently a more expressive representation of their contents. Thereafter, the image semantic concepts are modeled and classified by using a finite mixture models that help approximate the class distributions of the relevant salient objects. The resulting multi-level image annotation provides a more satisfactory semantic retrieval based on various keywords expressed at different semantic levels.

We can also mention the multi-platform Java desktop application '*AISO*' designed by Lingutla et al. [167]. Indeed, '*AISO*' tool extends the source code of the *Interactive Segmentation Tool*[1] (IST), originally developed for comparing the performance of image segmentation algorithms. '*AISO*' is an interactive tool that provides the researchers and curators the opportunity to work with two alternative modes of operation: in the first step, users delineate the portions of images into multiple highlighted segments through the *Interactive Graph Cuts* (IGC) [31]. In the second step, users annotate the resulting segments with a biological Ontology-based controlled vocabulary. The ontological terms are provided through the lightweight *Plant Ontology Web service* [51]. Moreover, the users may assign a taxonomic name to each entire annotated image by using the *uBio namebank search Web service*[2]. The quality of the segments of the annotated image can provide training data sets for developing applications of data mining, machine learning, predictive annotation, semantic inference, and comparative analysis.

The second type of semantic annotation methods pay more attention to the spatial positions of the objects depicted on the images as well as the spatial relationships that exist between them. Indeed, the use of spatial information helps enrich the semantic description of images and enhance the precision of the queries handled for an automated retrieval. Fig. 7 illustrates the progress of the semantic annotation process using the spatial relationships that exist between the image's objects.

For example, the image annotation tool suggested by Hollink et al. [106] provides the functionality to ensure an automatic region segmentation of an art painting image collection, which are then manually labelled with annotation concepts [105]. In fact, when a user decides that all relevant regions are labelled, the system carries out the calculation of the spatial information. Two types of spatial concepts are considered: (***i***) the absolute positions of objects (e.g., *east*, *west*) and (***ii***) the relative spatial relations between objects (e.g., *left*, *above*). Thus, existing ontology concepts are used to specify the positions as well as the spatial relations that are presented through an RDF schema. The spatial relations are extracted from SUMO [208], which is a large and well-structured ontology that takes into account Cohn's ideas about the

[1] http://kspace.cdvp.dcu.ie/public/interactive-segmentation/index.html
[2] http://www.ubio.org/index.php?pagename=xml_services

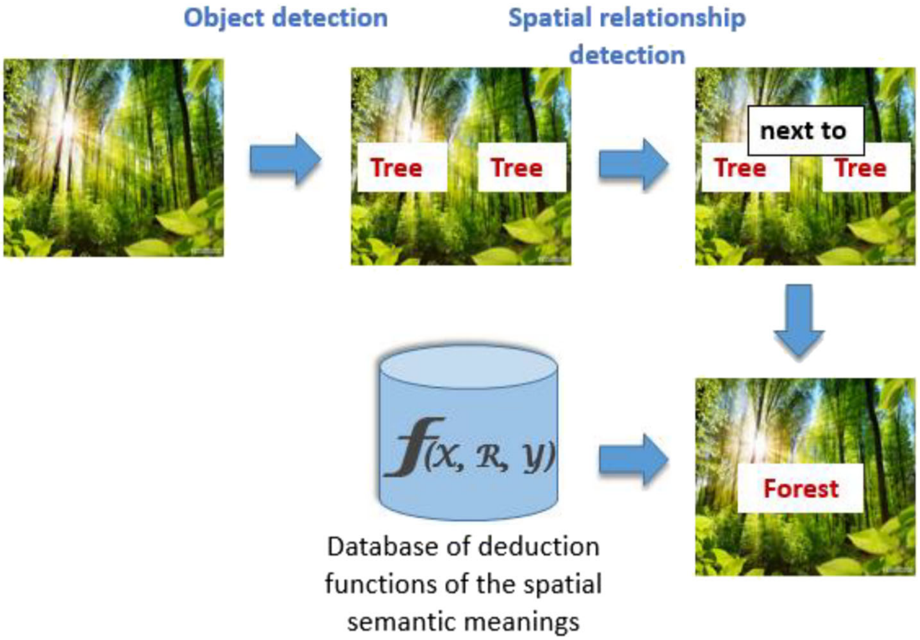
Object detection
Spatial relationship detection
Tree
Tree
next to
Tree
Tree
$f(x, \mathcal{R}, y)$
Database of deduction functions of the spatial semantic meanings
Forest

Fig. 7 Progress of the semantic annotation process using the spatial relationships that exist between the image's objects

spatial relations. The absolute positions are extracted from the general lexical of *WordNet* database.

Likewise, Muda et al. [200] used spatial information derived from an ontology. In this study, spatial annotations are addressed to enhance the precision of queries handled at the automated semantic retrieval. Indeed, assuming that a preliminary object segmentation and annotation step is realized (by adopting one method or another deriving from the literature), the design and implementation of this tool help automatically extract, identify and deliver the absolute spatial position of each object depicted on an image by providing the couple of coordinates (x_c, y_c) of its center of gravity (C). The resulting absolute spatial information help calculate and infer the relative spatial relationships of each pair of objects according to a set of predefined inference rules. A general heuristics helps infer 3-dimensional annotations, which indicate the relative closeness of the depicted objects to the viewer, by calculating the relative order of magnitude height information.

3.1.3 Summary

Despite the fact that the low-level feature-based image annotation methods help reduce the '*semantic gap'* problem by providing abstract-anatomical descriptions for the images, the results are often considered subjective and far from describing the semantic details.

As an alternative, semantic-oriented annotation methods are originated. These methods help generate and infer semantic descriptions, reflecting the image semantic content, either in terms of their visual characteristics (regions, objects, etc.) or in term of the spatial relationships existing between the objects that appear on the images. However, the expression of the image semantic content is considered partial. This is due to the absence of a predefined and precise semantic model (or pattern), which permits to define all semantic views and facets that must be described for providing a complete semantic description of the semantic content of such an image. In addition, even though visual content-based image annotation methods made it possible to reduce the problem of the '*semantic gap*', by adopting object recognition techniques and occasionally recognizing their semantic facets, the recognition of the activities and events that appear on the images is almost unnoticeable.

3.2 Users' tag-based image annotation methods

The collaborative aspect of the *World Wide Web* has given users the possibility to not only share images, but also to associate them with free textual descriptions known as keywords (or tags). However, these tags can be noisy, subjective, superfluous, ambiguous and/or missing, whence they are not considered as a reliable solution when it comes to image annotation. Therefore, numerous research studies have dedicated considerable efforts to refine the users' initial keywords. A refinement process helps certainly improve the quality of the image annotations. The policy consists in crushing the noisy and redundant labels, adding new-more expressive ones, finding a more-logical organization for the tags, etc. The annotation based on the refinement of the users' tags can be based on the calculation of the score of the different relationships existing between the keywords or the richness of structured vocabulary of external resources, such as ontologies.

3.2.1 Image annotation based on the measurement of semantic relationships between tags

Numerous Web portals, like *Flickr* and *Delicious*, offer users the ability to share images that they manually associate them with freely chosen tags [251, 275, 277]. However, the tagging information are often ineffective for indexing images. Therefore, it is necessary to adopt refinement techniques. These techniques can rely either on semantic similarity measurements [83, 199, 231, 342] or co-occurrence measurements [147, 316, 317]. Fig. 8 illustrates the progress of the semantic annotation process based on the calculation of the score of the semantic relationships between tags.

For example, Quattrone et al. [231], illustrate that the real-world folksonomies are characterized by power law distributions of tags and not commonly use similarity metrics. Indeed, *Jaccard coefficient* and *Cosine similarity* often fail to compute the semantic similarity between tags. As an alternative, a new metric is developed to capture the semantic similarities between the large-scale folksonomies in order to increase the accuracy of the researches realized on the images. This metrics is based on a mutual reinforcement policy: two tags are deemed similar if they are associated with similar resources, and vice-versa two resources are deemed similar if they are labelled with similar tags.

The research study of Mousselly-Sergieh et al. [199] consists in classifying the concepts associated with the images obtained in real-life from specific geographical locations, and shared later on the *Flickr* social network. Two steps are defined: the first one consists in identifying and distributing the tags that frequently occur in the folksonomy on a set of clusters. Thereafter, the probability distributions of each other tag is derived based on its co-occurrences with the most frequent tags in the folksonomy. After computing the different probability distributions, the distance between each pair of tags is calculated according to the extended *Jensen-Shannon Divergence* measure (JSD), known as *Adaptation of the JSD*

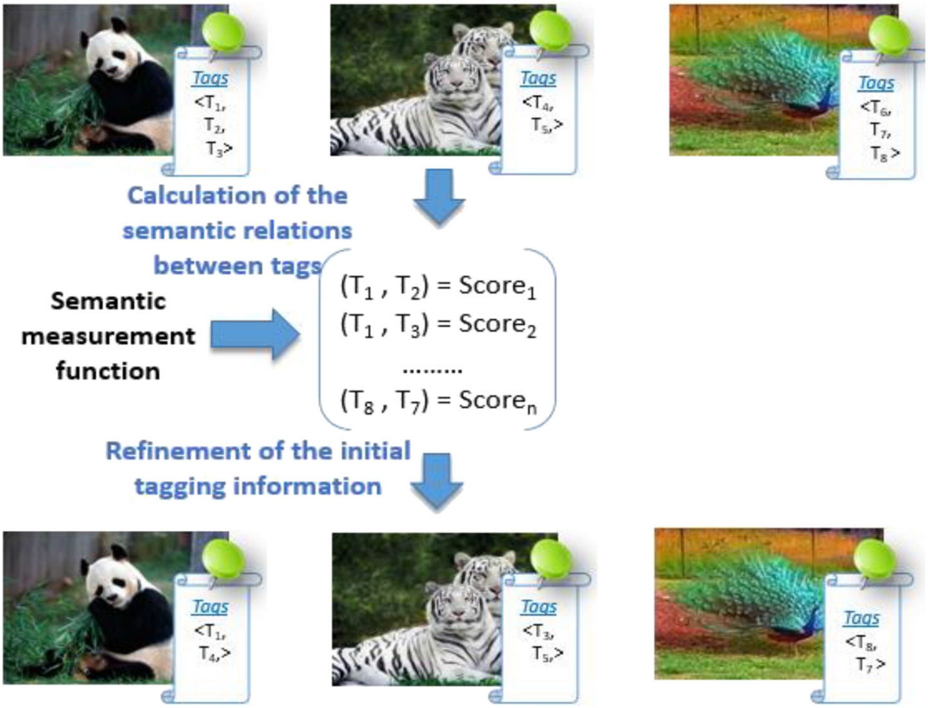

Fig. 8 Progress of the semantic annotation process based on the measurement of semantic relationships between tags

measure (AJSD) [187], such as: two tags are considered similar if the distance between their distribution is under a certain threshold. Therefore, the resulting classification of tags helps improve the quality of the image annotations by adding new similar concepts to the initial concepts.

Ksibi et al. [147] proposed to improve the quality of browsing results provided by *Flickr* social search engine. Indeed, a weighted graph is first constructed by using a new-suggested measure named *Second Order Context Flickr Similarity* (SOCFS). This graph helps represent the different co-occurrence relationships, which exist between the social data. The initial tags are then refined by estimating the set of relevant concepts based on the weighted graph.

Xu et al. [317] measured the semantic relatedness between two images shared on *Flickr* social network, according to their associated tags, in order to improve the clustering and searching processes. The proposed approach follows a set of steps: in the first one, four functions are defined based on the information theory to measure the semantic relatedness between the tags. The second step consists in integrating each pair of tags on a bipartite graph to remove noisiness and redundancy that they cause. The final step helps add the order information of the semantic relatedness of the tags so that the tags with higher positions become favorited.

3.2.2 Image annotation based on the richness of structured knowledge resources

The policy of this type of annotation method consists in refining the socio-tagging information associated with the images thanks to the richness of external semantic resources in order to

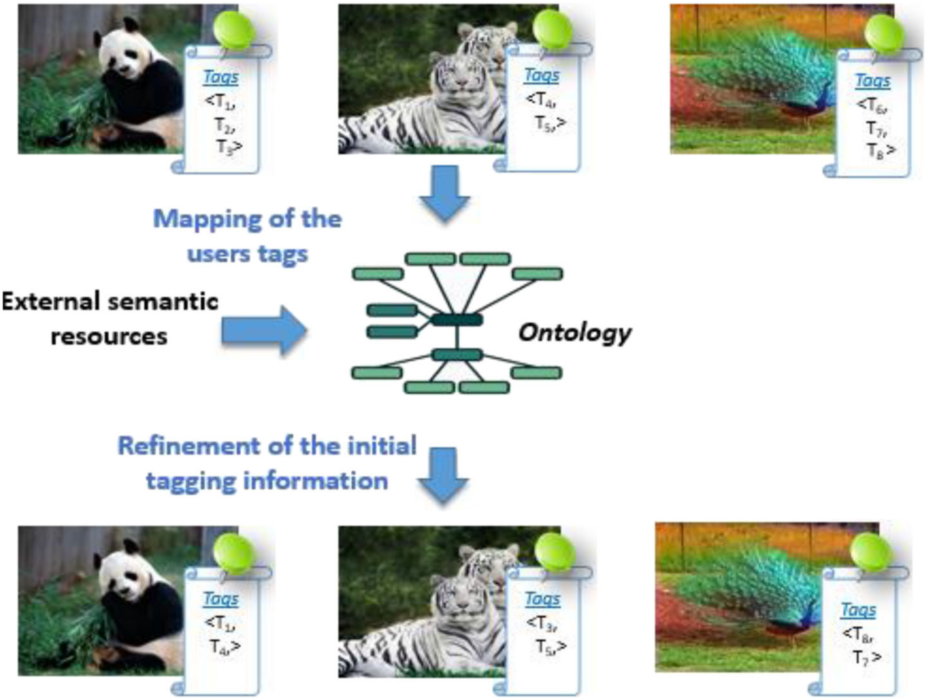
Tags
<T1, T2, T3>
Tags
<T4, T5>
Tags
<T6, T7, T8>
Mapping of the users tags
External semantic resources
Ontology
Refinement of the initial tagging information
Tags
<T1, T4>
Tags
<T3, T5>
Tags
<T8, T7>

Fig. 9 Progress of the semantic annotation process based on the richness of structured vocabularies

generate more pertinent annotations [27, 112, 113, 120, 125, 170]. The accredited refinements can have different forms: adding new semantic information, removing the noisy information, reordering the tagging information according to a priority function, etc. Fig. 9 illustrates the progress of the semantic annotation process based on the richness of structured vocabularies.

In this context, we can mention the '*STAG*' and '*Linked tag*' systems, designed by Im et al. [112] and Im et al. [113], respectively. Indeed, '*STAG*' system aims at providing the semantic relationships connecting the users' tags by using *Dbpedia*[3]. The final annotations associated to the images are presented in the form of triplets, namely (*Tag, Relationship, Tag*), for a later SPARQL interrogation. In '*Linked tag*' system, the authors have added a new method for tag-ranking that exploits the RDF annotation graph. More specifically, the most relevant tag is placed at the first position and the least relevant is placed at the last one.

'*iTagRanke*r' is also a socio-tagged image annotation system designed by Jeong et al. [120]. It consists first in collecting the tags from similar images, and then propagating them towards the untagged image in question. Thereafter, the collected tags are reorganized according to their semantic relevance versus the image. The reorganization of the tags is based on a matrix that contains the semantic relatedness measures of all tags by favoring labels over others. The computing of the semantic relatedness degrees between each pair of concepts is performed based on the *Lcn* (*Leacock et al.*) [157] and *Lin* [166] measurements and by relying on the *WordNet* Ontology.

'*Tourism-Annotizer*' tool [27] helps annotate 25.000 images derived from *Flickr* social network. The first step consists in modeling the semantic content of the images by using an *Extended Conceptual Graph Formalism* (ECGFs) and relying on the '*Touring Ontology*'. The

[3] http:// www.dbpedia.org/

output of this step consists of a pattern that helps model the different semantic facets of an image, namely: the taxonomic, temporal, spatial and qualificative facets of the visualized objects, unfolding events and activities. The second step consists in building a set of semantic rules, which are characterized by their types. These rules are expressed by the logics of predicates. The application of any semantic rule generates a semantic value that is associated with the image. An inference engine is implemented to create new semantic rules, and consequently to generate new semantic results.

3.2.3 Summary

The image annotation method based on the refinement of the tags associated with the images shared on the social networks is sometimes based on the calculation of the semantic similarities between tags, and sometimes on their co-occurrence relationships. This method of annotation is purely automatic, so it can be applied to annotate large collections of images. However, it is not considered an effective solution to express the semantic content of the images because it is based on static calculations not founded on a logical reasoning.

As an alternative, a method that consists in automatically ensuring the refinement of the socio-tagging information based on structured-external resources emerged. This method helps generate more pertinent annotations expressing not only the objects depicted on the images but also the unfolding events and activities. However, the expression of the semantic content of images is often considered a delicate task due to the need of a predefined-semantic model or pattern.

4 Visual content-based image annotation techniques

In the field of digital communication, images are represented using low-level features, namely: texture, color, shape, etc. Since an image is an unstructured array of small integers called pixels, the main step in the semantic understanding of its content is to extract the effective and efficient visual features from these pixels. Indeed, an appropriate visual feature representation helps significantly improve the quality of the semantic learning and annotation results. It should be mentioned that the visual feature extraction is preceded by a global or region-based representation. In the approaches based on a global representation, a global feature vector is extracted from the whole image, such as color correlogram, color histogram, edge direction histogram, and so on. The global features are useful for classifying simple scene categories, like '*mountain*', '*sunset*', '*building*', etc. The regional approaches require prior image segmentation followed by a visual feature vector representation of each generated region. In general, users are often more interested in specific regions rather than in the entire image given that the image representation based on a regional level is proved to be closer to the human perception system [127]. In this paper, we focus on the *Region-based Image Representation* (RBIR), which is the cornerstone for many image annotation and research approaches. Specifically, image segmentation, feature extraction and semantic learning algorithms and techniques are reviewed successively.

4.1 Image segmentation

Image segmentation plays a significant role in computer vision, object recognition, tracking, and image analysis as a preprocessing step [164]. There is a great amount of literature on segmentation, among which region-based methods have shown promising results [133]. Indeed, for the region-based image representation, a fundamental step is to divide the images into multiple components according to the homogeneity of their visual features. Therefore, image segmentation techniques and algorithms have been the focus of numerous image annotation and research studies in the literature. This will be the subject of this sub-section.

The grid-based image segmentation is a simple segmentation technique that consists in breaking down the image into blocks [13, 190, 198, 228, 69]. The visual features are then extracted from these blocks. Although the block-based segmentation is not greedy in terms of computation, it still has a limitation for multi-object problems in a segmented region. In addition, region features are usually not accurate because it is difficult to determine the size of blocks for the image representation. It is so recommended to use this segmentation technique in domain-specific applications, such as medical image archival and analysis [98].

The edge-based image segmentation technique consists in evolving a segmentation curve around an object [35, 36, 151, 340]. For instance, in order to segment an object depicted on an image, the active contour curve is evolving from an initial point and stopping when it coincides with the boundary of this object. Therefore, the choice of the starting point of the curve is the main issue of this technique. Some existing methods have proposed to manually assign the initial seed point to start active contour, and others perform a reset if the first initialization did not return the correct boundary of the object, which is an expensive task in terms of computation. The edge-based image segmentation technique is generally used in specific domains, such as image preprocessing applications [3]. For more details on the edge detection techniques readers can refer to [145].

The clustering algorithm-based image segmentation technique, like *k-means* and *Fuzzy k-means* algorithms [203], work usually as follows: At first, an image is divided into a set of blocks of size (4x4) pixels. For each block, the texture and/or color features are extracted. A clustering algorithm is then executed in order to cluster the feature vectors of the blocks. Therefore, the pixels that belong to the blocks of the same cluster constitute a region [5, 150, 194, 282, 302]. The main problem of this segmentation technique is the need to predefine the number of segments. In addition, an unsuitable choice of the number of clusters *k* can generate poor results. The choice of the optimal centroids is also a complex task.

The statistical model-based segmentation technique assumes that the image's objects are understood by a certain pattern. The list of the models frequently used for image segmentation are: *Object Background/ Threshold Model* [248], *Markov Random Field Model* [60, 76, 205, 214], *Neural Model* [296], *Fractal Model* [222], *Fuzzy Model* [33], *Multi-resolution* [271] and *Transformation model*, namely: *Watershed model* [295] and *Wavelet model* [37]. For a deep studying, interested readers can consult [11].

The graph-based segmentation technique consists in modeling an image with a weighted-undirected graph. Indeed, a pixel is associated with nodes and edge weights that define the (dis)similarity between the neighboring pixels. The image's graph is then partitioned according to a criterion designed to model the best clusters. Each resulting partition is considered as a segment of the image. The more popular graph-based segmentation algorithms are: *Random Walker* [86], *Normalized Cuts* [254], *Minimum Spanning Tree-based Segmentation* [330], *Isoperimetric Partitioning* [87] and *Segmentation-based Object Categorization*. The shortcoming of this segmentation technique is that the finding of the optimal partition is a computationally expensive task.

The region-growing method is a typical serial region segmentation algorithm [132]. Its main idea relies on the assumption that the neighboring pixels within one region have similar values. The usual procedure requires first to select a seed pixel, and then compare it with its neighbors. If a similarity criterion is satisfied, the pixel will be merged in the cluster as one of its neighbors. Indeed, the regional growth algorithm is simple and requires only a few seed points to be executed. It helps separate the connected regions with the same characteristics, and usually provide good boundary information and significant segmentation results. The selection of the similarity criterion can be freely specified even a large number. The disadvantage of this algorithm is that the computational cost is significant [10]. Besides, the noise and grayscale unevenness can lead to emptiness and over-division. The shadow effect on the image is also not very good [289].

Deep convolutional neural networks for semantic image segmentation have recently become a dynamic study field [308]. Indeed, since the large breakthrough in deep learning, the research efforts have been oriented towards the CNN-based approaches [165]. The relentless success of deep learning techniques in various high-level computer vision tasks, like the supervised approaches of CNN for image classification and object detection [146, 260, 270], has motivated researchers to explore the effectiveness of such networks for pixel-level labelling problems, such as semantic segmentation. Therefore, many segmentation techniques have been suggested, among which we can mention: *Fully Convolutional Network* [174], *SegNet* [17], *Bayesian SegNet* [139], *DeepLab* [45, 46], *MINC-CNN* [20], *CRFasRNN* [338], *DeepMask* [226], etc. Readers can refer to [75] for more detail on the semantic segmentation methods based on deep learning.

4.2 Feature extraction

As previously mentioned, the first step at the *region-based* image representation is the image segmentation. At this stage, the visual features are extracted from the segmented regions to be later on the focus of the semantic learning, classification and annotation processes. This section includes a brief description on the different low-level feature extraction, feature descriptors and deeper feature extraction. Available extraction algorithms are also illustrated.

4.2.1 Low-level feature extraction

Indeed, the image annotation and retrieval systems often require an analysis of the image content, which might refer to the color, texture, shape and spatial relationships of the image segments.

- **Color features** Color features are among the most important and wide used components in image retrieval systems. Color features are defined according to particular color space or model, such as RGB, LAB, LUV, HSV (HSL), YCrCb and the hue-min-max-difference (HMMD) [168, 185, 264]. After specifying the color space, the color features can be extracted from the segmented regions. Numerous color features have been suggested in the literatures, such as *color moments* (CM) [111, 118], *color histogram* [117, 269], *color correlogram* [110], *color coherence vector* (CCV) [217], etc. On his side, MPEG-7 [186] standardizes various color features, like *color layout descriptor* (CLD), *dominant color descriptor* (DCD), *scalable color descriptor* (SCD) and *color structure descriptor* (CSD). Table 2 [336] presents a summary data on the different color descriptors with their advantages and disadvantages.

Table 2 Contrast of different color descriptors

Color method	Pros	Cons
Histogram	Intuitive, simple to calculate.	Important dimension, susceptible to noise, ignoring spatial information.
CM	Vigorous, compact.	Does not describe all colors, ignoring spatial information.
CCV	Considering spatial information.	Expensive computational cost, important dimension.
Correlogram	Considering spatial information.	Expensive computational cost, ignoring scale and rotation, susceptible to noise.
DCD	Vigorous, compact, perceptual meaning, considering spatial information.	Need for post-processing for spatial information.
CSD	Considering spatial information.	Susceptible to noise, ignoring scale and rotation.
SCD	Scalability, compact on need.	Less precise if compact, ignoring spatial information.

• **Texture features** It is usually thought that image annotation and retrieval systems use texture features for recognition and interpretation. Thus, a large number of techniques have been suggested to extract this type of features. According to the domain from which the texture features are extracted, we can distinguish into spatial and spectral texture feature extraction methods.

In essence, the spatial texture feature extraction approach consists in calculating the statistics of pixels or finding the local pixel structures within the original image domain. The spatial texture feature extraction techniques may be categorized as structural, statistical and model-based, with:

The structural techniques [96, 159, 191] represent the texture by well-defined primitives (*micro-texture*) and a hierarchy of spatial arrangements (*macro-texture*) of those primitives.

The statistical techniques represent the texture by using non-deterministic properties that govern the distributions and relationships between the grey levels of an image. *Tamura texture features* [123, 273, 99], *Moments* [216] and features derived from *Grey Level Co-occurrence Matrix* [173] are the common statistical features of the spatial domain.

The model-based texture analysis uses stochastic or generative models. Indeed, the underlying texture property of an image is characterized by model parameters. *Fractal Dimension* (FD) [40] and *Markov Random Field* (MRF) [52, 286] are from the most popular texture models.

In spectral texture feature extraction approach, an image is represented in a space whose coordinate system has an interpretation, which is closely related to the texture features (such as size or frequency). *Fourier Transform* (FT) [239], *Gabor filters* [30, 57, 21, 184], *Discrete Cosine Transform* and *Wavelet Transforms* [111, 153, 178, 183] are the common techniques of the spectral domain.

Table 3 [225] provides a summary data on the different texture methods with their advantages and disadvantages.

• **Shape features** Shape features help human beings identify and recognize the real-world objects. Thus, this type of features was frequently used in numerous image annotation and retrieval applications. Shape feature extraction techniques are classified into two main categories [332]: *contour-based* [70, 187, 327] and *region-based* [65, 167, 227] techniques. Indeed, *contour-based* techniques compute shape features by using only a portion of the region that is the boundary of the shape. Therefore, they are more susceptible to noise than

Table 3 Contrast of texture features

Texture method	Pros	Cons
Spatial texture	Easy to understand, meaningful, can be extracted from any shape without losing information.	Important dimension, susceptible to noise, ignoring spatial information.
Spectral texture	Need less computation, vigorous.	Need square image regions with sufficient size, ignoring semantic meanings.

region-based techniques. *Region-based* techniques help extract shape features from the entire regions. They are frequently employed by colored images annotation and search systems. Yang et al. [322] presents a survey on the existing shape-based feature extraction approaches. Fig. 10 shows the classification hierarchy of the shape feature extraction techniques [322].

- **Spatial relationships** Spatial relationships are also considered in image processing. They help reveal the locations of objects depicted on a given image as well as the spatial relationships

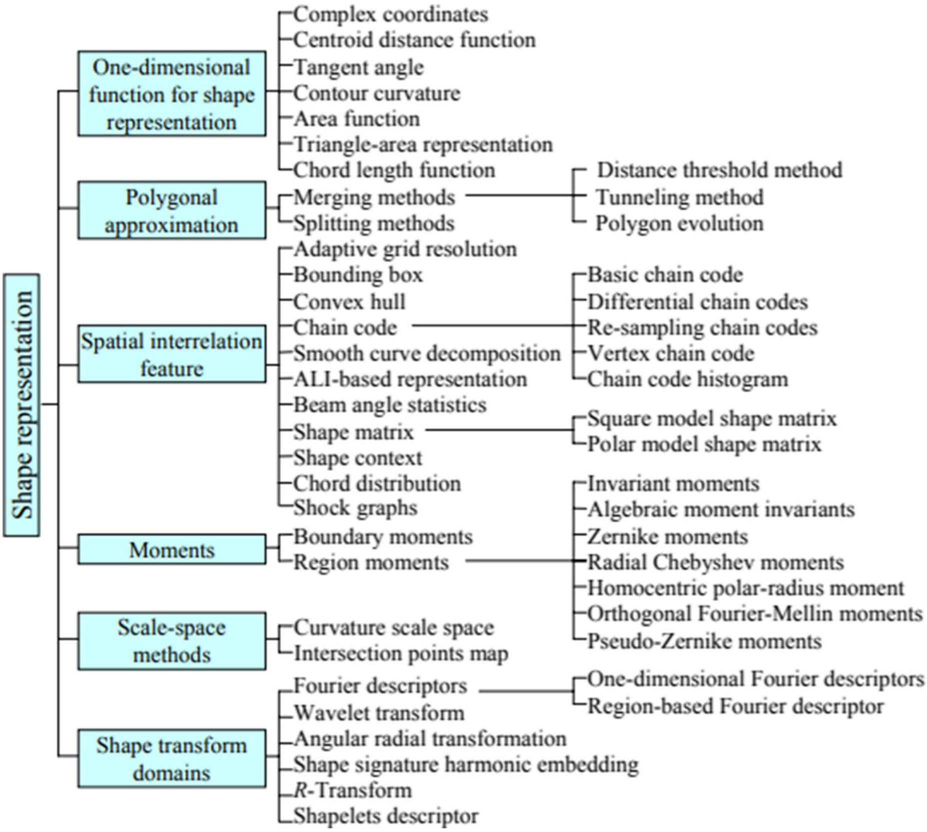

Fig. 10 An overview on the shape description techniques

between the different objects. We can mention two main cases: absolute spatial location of regions (e.g., *east*, *west*) [106] and relative location of regions (e.g., *left*, *above*) [16, 106].

4.2.2 Feature descriptors

Over the last two decades, some descriptors, like SIFT and SURF, have been used on object recognition for image annotation and retrieval purposes. These descriptors are defined by some invariance properties, like *scale*, *rotation*, *viewpoint* and *illumination*. It should be mentioned that other descriptors of scene recognition, like GIST, bypass the segmentation and the processing of individual objects or regions.

- **SIFT features** The *Scale Invariant Feature Transform* (SIFT) was proposed by Lowe et al. [177]. The policy of SIFT consists in estimating key points' location, orientation and scale to create the descriptor. The SIFT descriptor was founded on the human vision behavior. It privileges gradients and orientations to slightly move location in order to recognize objects. It computes gradients on key point's region that is partitioned into (4×4) sub-regions. From each region, the orientation histograms are estimated. SIFT performs better on invariant rotations and scale changes but neither in the case of low-contrast and illumination changes within an image. Some years later, the SIFT descriptor was normalized by $L2 - norm,$ inducing illumination invariances. The rotation and scale invariances are calculated from detector's information. The SIFT descriptor was employed in many TBIR systems, like [25, 62, 323].

- **SURF features** The *Speeded-Up Robust Feature* (SURF) was introduced by Bay et al. [18]. It needs key points' location and scale to create the descriptor. The descriptor estimates first interest point's orientation and then the gradients' approximation (d_x and d_y). The key point's region is partitioned into sub-regions, where for each one, $\sum d_x$, $\sum d_y$, $\sum |d_x|$ and $\sum |d_y|$ are computed. SURF descriptor is also normalized by $L2 - norm$. It is a sparse, scale and rotation-invariant descriptor, which performs better in the case of repeatability, robustness and distinctiveness. It is also robust to noise, geometric, detection errors and photometric deformations. SURF performs better at low illumination within an image. Recent TBIR systems have used SURF descriptor, like [258, 312].

- **GIST features** GIST descriptor is a computational model, proposed by Olive et al. [212], which focuses on the shape of the scene itself, on the relationship between the surface outlines and their properties, and neglects the objects depicted on the images and their relationships. The policy of GIST descriptor is based on a very low dimensional representation of a given image, which is called Spatial Envelope. A set of perceptual dimensions (*openness*, *naturalness*, *expansion*, *roughness* and *ruggedness*) are proposed in GIST to represent the dominant spatial structure of a given scene. These dimensions can be effectively estimated by using the spectral and coarsely localized information. The proposed model generates a multi-dimensional space in which scenes that share membership in semantic categories, such as highways, streets, and coasts, are projected closed together. The performance of the Spatial Envelope model demonstrates that specific information about object shape or identity are not necessary for scene categorization and the modeling holistic representation of the scene generates information about its probable semantic category. The GIST descriptor has been used in TBIR systems, like [281, 323].

4.2.3 Deeper features

Because of the diversity of appearances, backgrounds, rotation of cameras, object scales and illumination conditions, it is hard to manually design an efficient feature descriptor that helps describe all object types. With the rapid advancement in deep learning, deep *Conventional Neural Network* (CNN), which can learn deeper features, was introduced in order to address the different issues caused by the traditional architectures. The CNN model is investigated in Sect.4.3.2.

4.3 Semantic learning

The image feature extraction is followed by a higher-level semantic learning. In the beginning, the image semantic learning was based on the use of relevance feedback technique (RF) [241, 313]. Nevertheless, this type of learning causes similar issues than the traditional manual annotation approach. As an alternative, automatic image annotation approaches (AIA), using *machine learning* techniques (ML), have emerged.

4.3.1 Machine learning

The increasing volume of digital images in numerous fields, the availability of different types of data and the progress of computational processing have made ML an important aspect in the *Artificial Intelligence* (AI). Indeed, ML techniques are intended to enhance the learning competence of computers and construct models that help predict future data. ML is the most important data analysis method, which iteratively learns from the available data by using algorithms. The Iterative follow-up is performed thanks to models, which are programmed to accept new data. Significant predictions and decisions may be provided by these models. The ML techniques are classified into two distinct types, namely: *supervised-learning* (SL) and *unsupervised-learning* (UL). Recently, *deep-learning* techniques (DL) emerged as new instances of ML. Fig. 11 shows the diagrammatic representation of ML techniques.

• **Supervised-learning** *Supervised-learning* (SL) is the research of algorithms, which reason from externally supplied instances in order to produce general hypotheses that constitute predictions about future instances. In other words, the aim of the SL is building a concise model of the class labels distribution in terms of predictor features. The resulting classifier is thereafter used to assign class labels to the testing instances, where the predictor feature values are known but the class label value is unknown. Fig. 12 illustrates gradually the process of SL.

SL is the most used technique in applications where available past data predict the expected future events. Equation (1) shows the general representation of SL [211] as:

$$D = \left\{ (X_i, y_i)_{i=1}^{N}, X_i \left(x_i^1, \ldots\ldots, x_i^d \right) \right\} \tag{1}$$

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

- The *k-Nearest Neighbor* (*k*-NN), which is known as a simple and efficient approach, is a non-parametric supervised classifier. It has been adopted since the early 1970's in statistical applications [71]. Assuming that a distance function is used (e.g., *Euclidean*

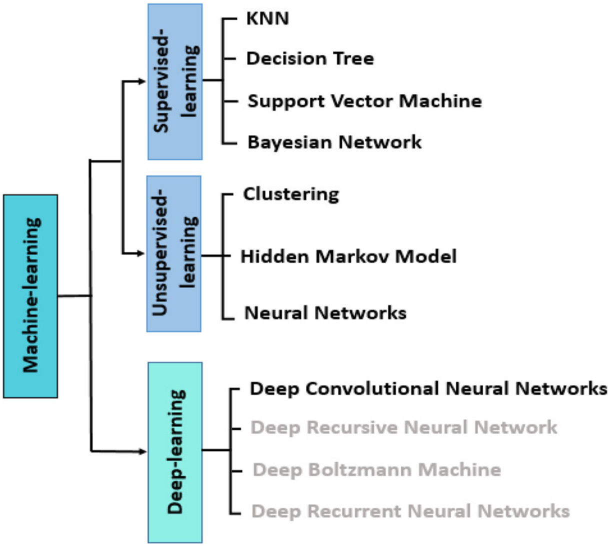

Fig. 11 Diagrammatic representation of SL techniques

distance and *Manhattan distance formula*), the primary theory behind *k*-NN consists in finding a group of *k* samples in the calibration dataset, which are nearest to unknown samples. From these *k* samples, the label (or class) of unknown samples are determined by computing the average of the response variables (i.e., the class attributes of the *k*-nearest neighbor) [6, 309, 229]. *K*-NN was the focus of many research studies. For instance, an algorithmic model for automatic classification of different types of flowers using *k*-NN classifier was proposed by Guru et al. [91]. This model is based on textual feature extraction preceded by a threshold segmentation method. In Ref. [131], ten morphological characteristics have been analyzed to identify four Monogenean species of fishes. An accuracy of 91.25% was yield using a *k*-NN classifier. Ref. [121] proposes a new *Visual-Semantic Nearest Neighbor* (VS-KNN) method by collectively exploring visual and semantic similarities for image

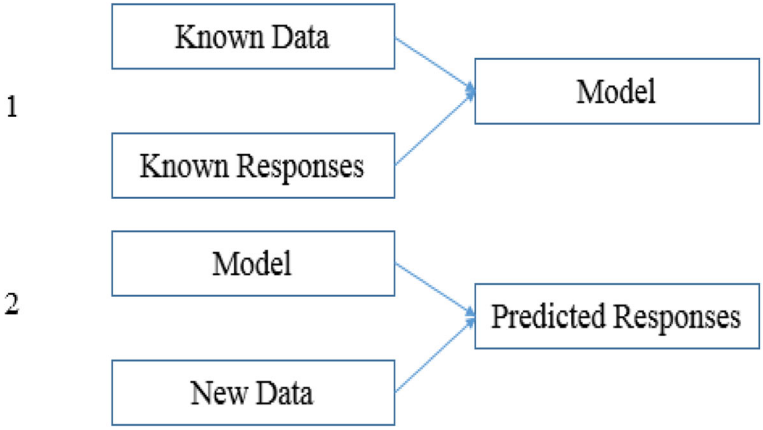

Fig. 12 Steps of the supervised learning process

annotation. In Ref. [219], a categorization based approach is presented for an automatic image annotation. Images are first segmented using *k*-means clustering, and then processed to form color and texture feature vectors. The feature vectors are tested using *k*-NN. The system is validated using ten categories from COREL images.

- The *Decision Trees* (DT) are multi-stage decision tools [232, 233, 32]. They are named binary or n-ary trees according to the number of decisions taken at each of their internal node. DTs accept inputs as a situation described by a set of attributes and provide the predicted output for the given input later on. The input/output relationships may be expressed by using human understandable rules, such as *if-then* rules. The DTs are trained by using a group of labelled training samples that are characterized by a set of attributes. The policy of construction of a DT consists in recursively partitioning the training samples into groups without overlapping. At each division step, the used attribute is ignored. The process continues until all samples of a group are put in the same class or when the tree achieves its maximum depth and any attribute persists to divide them. In order to label new samples, a tree is crossed from the root node towards a leaf node by using the attribute values of the new samples. It should be mentioned that several DT induction methods have been proposed in the literature, such as ID3 [232], C4.5 [233] (improved version of ID3), and CART [22]. Just as illustration, Sethi et al. [250] adopted the CART methodology for classifying outdoor images into four classes, namely: '*marine*', '*sunset*', '*nocturne*' and '*arid*' images. For image representation, each component of HSL color space is partitioned into 8 intervals. The 24 obtained intervals (3x8) are used as attributes for images. In the study of Wong et al. [313], image acquisition parameters, such as *exposure time*, *aperture* and *focal length*, are considered as attributes during the scenery images annotation and classification processes. The C4.5 method is used to learn decision rules for mapping these attributes to the image semantics. In the study of Shyu et al. [259], a C4.5 decision tree is built based on a set of images that are relevant to the query. This tree is then used as a model to classify database images into relevant and irrelevant classes. A similar methodology is employed by Low et al. [176] to enhance the performance of relevance feedback of image retrieval systems. ID3 decision tree is designed to classify the images as relevant or irrelevant based on their color features. Recently, a classifier using DT and *Rough Sets* (RS) [221] is designed by Patil et al. [220] to tag untagged images. The suggested method helps accurately classify the images of the database by using the strength of *texture and color image features*. Tallapragada et al. [272] proposed a semi-decision algorithm that can target only the tumor parts from the medical images. The identification and morphological processing of the tumor images are based on a thresholding segmentation. Thereafter, texture-based techniques are used to extract the feature vectors from the segmented regions. Finally, medical images under-test are classified by using a DT classifier.
- The *Support Vector Machine* (SVM) is one of the standard ML algorithms that follows statistical learning methods. Thanks to powerful theoretical foundations available, SVM classifier has been frequently used to learn top-level concepts from low-level image features. SVM helps classify linear and non-linear data by using kernel mapping. It helps reach the optimal class boundaries by determining the maximum distance between the classes. From a training dataset, an SVM classifier works by dividing the data into distinct

sets via an optimal separator named hyper-plan. Therefore, the main objective of SVM is to find the hyper-plan that separates the points closest to the separator in a data space. The data points nearest to the separator are labelled as support vectors. An SVM is essentially a binary classifier. Assume an example where it exists a set of training data $\{x_1, x_2,...,x_n\}$ as vectors in space $X \subseteq S$ belonging to two distinct labeled classes $\{l_1, l_2,...,l_n\}$, with $l_i \in \{-1, 1\}$. The aim is to find a hyper-plane to separate the data. Indeed, it is possible to define many hyper-planes. However, the optimal separating plane is the one which helps maximize the margin between the nearest data point of each class and the hyper-plane. Since that automatic image classification and annotation require multiclass classifiers, it is necessary to train a distinct SVM for each concept, where each SVM provides a probability value. The final class label attributed to the image is generated by merging the output decisions of the all classifiers. SVM has been frequently used to resolve divers classification issues, such as object recognition, text classification and image annotation [7, 29, 38, 42, 53, 68, 74, 81, 94, 109, 123, 158, 161, 179, 228, 230, 249, 255, 278, 285, 307, 321, 337, 343]. For example, Feng et al. [68] addressed the problem of providing large labelled training data needed in the training step of a classifier to annotate the big collections of images. The main idea consists in starting from a small set of labelled training images, and successively annotate a larger set of unlabeled images by using the co-training approach. Two statistically independent classifiers are used to co-train and co-annotate the unlabeled images. Thus, the bootstrapping approach has been used to co-train different features (color histogram, texture and shape) extracted from two segmentation methods by using SVMs. An effective framework for AIA is suggested by Huang et al. [109]. It consists in dividing respectively the main objects and background objects from an image, and then extracting their color, texture and shape features. Indeed, a combination of *Active Contour Model* (ACM) [137] and JSEG algorithm [58] is leveraged to segment and detect the main objects in an image. The main classifier and background models are trained using the object-based feature vectors. *Gaussian Mixture Model* (GMM) is employed to explore the relationships between image classes and image backgrounds based on the built association knowledge base. Wei et al. [307] addressed the problem of traditional methods providing poor experiment results due to the learning of the co-occurrence of keywords and images, and the ignorance of the correlation between keywords. Indeed, an automatic image annotation approach, which helps reach a higher accuracy by using multi-class SVM with ontology, was proposed. Specifically, semantic dictionary *WordNet* is used to calculate the correlations between keywords of the derived hierarchy. To present the image visual features and apply a mixed kernel in multi-class SVM, Bags of Visual Words model is employed. Finally, the probability outputs from multi-class SVM and the word correlations probability calculated from ontology are combined to provide a final result. More recently, Alham et al. [7] introduced a distributed SVM algorithm for large-scale image annotation (MRSVM), which divides the training dataset into smaller subsets and trains SVM in parallel by using Map-Reduce pattern.

- The *Naive Bayes* (NB) classifier [63, 129, 331], is frequently employed in AIA approaches since it is a simple probabilistic classifier that is based on the Bayes theorem with strong assumptions. This classifier consists in building a probability model independent of features. Indeed, the *Naive Bayes* classifier supposes that the presence (or absence) of a particular feature of a class is unbound to the presence (or absence) of any other feature, given the variable of class. In essence, NB classifier helps annotate images with multiple semantic categories. Indeed, according to certain features extracted from an image, NB classifier determines the posterior probability that this image belongs to any particular category. Therefore, the image can be assigned to multiple categories. The images with the same category can be classified according to the probabilities. Rui et al. [243] proposed a new approach for auto image annotation. In the learning stage, image segments are grouped into region clusters using *k*-means algorithm with pair-wise constraints [299]. In the annotation stage, a *semi-naïve Bayesian* model (SNB) is employed to compute the posterior probability of concepts given the independent subsets of region clusters. Yavlinsky et al. [326] used the Gaussian and EMD [240] kernels to estimate the feature distribution. They used color and texture features for image representation if it consists of the Gaussian kernel and region-based image representation when it consists of the EMD kernel. Indeed, the regions are segmented by using a simple *k*-means clustering. The average of the kernel functions is measured in order to compute the conditional probability $p\ (I \mid c)$ for each image *I*. In the study of Sami et al. [246], an automatic image annotation approach, which integrates the *Naive Bayes* classifier with the *Particle Swarm Optimization* algorithm (PSO) [50, 140, 284] for classes probabilities weighting, is suggested. This hybrid approach consists in refining the output of multiclass classification that is based on the usage of *Naive Bayes* classifier to automatically label images with a number of words. Indeed, each input image is segmented by using the normalized cuts segmentation algorithm in order to create a descriptor for each segment. One *Naive Bayes* classifier is trained for each class. PSO algorithm is employed as a search strategy to identify an optimal weighting for classes probabilities from *Naive Bayes* classifier.

- **Unsupervised-learning** Unlike supervised-learning, in which the presence of the outcome variables guide the learning process, the learning dataset in the unsupervised-learning consists only of input vectors of unlabeled data. Indeed, the unsupervised-learning algorithms analyze the set of input data, group the data points based on perceived similarities and derive conclusions from these similarities. The most commonly used unsupervised-learning techniques are *clustering*, *Hidden Markov Model* (HMM) and *Artificial Neural Networks* (NNs). These techniques help explore the unlabeled data in order to identify intrinsic or hidden patterns.

- *Clustering* [14, 124] is an unsupervised-learning process, where one seeks to identify a finite set of categories for describing the items from a dataset. The groups generated following the execution of a clustering process are called clusters. Unlike supervised-classification that analyses class-labeled instances, clustering process has no training stage and it is often used when the clusters are not known in advance. Indeed, the attributes providing the best clustering should be often identified in a first stage. A similarity metric is then defined between objects of data so that similar data objects are grouped into the same cluster, while the different data objects are distant towards other ones. The clustering of data is based on the principle of maximizing the intra-cluster similarity and minimizing the inter-cluster similarity. A good clustering method helps produce high quality clusters with high-intra-cluster similarity and low-inter-cluster similarity. The efficiency of the clustering techniques depends on the use of algorithms as well as the functions for distance calculation. The quality of a given clustering method is also computed according to its ability to discover some or all of hidden patterns. The most common clustering techniques are the partitioning clustering and the hierarchical clustering [89, 136, 334]. In the study of Wang et al. [302], images are segmented into blocks of size (4x4) from which color, texture, shape, and location are extracted. Thereafter, *k-means* clustering is applied for grouping the feature vectors into several clusters with each cluster corresponding to one region. The clusters representing high-level categories, such as '*textured/non-textured*', '*indoor/outdoor*', '*objectionable/benign*' and '*graph/photograph*' help improve the retrieval process by narrowing down the searching range in the databases and ensuring semantically adaptive searching methods. A clustering algorithm is also proposed by Pandey et al. [215] for achieving a dataset with images grouped semantically. The resulting image dataset can be used in CBIR systems. Indeed, the visual feature extraction is preceded by a global representation. A combination of color histograms and moments, Gabor texture, and pseudo Zernike moments is adopted to provide vectors in color, texture, and shape feature spaces, respectively. The used clustering algorithm is based on the agglomerative method of hierarchical clustering algorithm. The used similarity measures are vector *cosine distance* for histograms, *L2 distance* for texture and shape, and weighted *L1 distance* for color moments. Kumar et al. [263] introduced an approach for image feature vector classification using an unsupervised clustering technique. The suggested approach aims at partitioning the trained image feature vectors into highly relative clusters. It consists of two stages : (*i*) Image pre-processing stage, and (*ii*) Classification stage. In the pre-processing stage, the set of the image feature vectors is trained from the set of grayscale images through the spatial-statistical operators. It consists of two steps: feature extraction and feature selection. In the feature extraction step, the input image is decomposed into (8x8) blocks. On each block, three spatial statistical operators are applied and three features from each individual block, such as average, standard deviation and variance, are extracted. In the classification stage, the trained image feature vectors are partitioned into "*m*" highly relative clusters using *k-means* algorithm and *Euclidean distance* measure.

- The *Hidden Markov Model* (HMM) is a finite state machine that has some fixed number of states. It permits to provide a probabilistic framework for modeling time series of multivariate observations. It consists of a statistical *Markov* model where the system being modeled is supposed to be a *Markov* process with unobserved (hidden) states. A HMM can be considered as a simplest dynamic *Bayesian* network. In HMM, the state is not directly visible, but the output, which dependents on the state, is visible. Each state has a probability distribution on the possible output tokens. Then, the sequence of tokens generated by an HMM provides certain information about the sequence of states. The general architecture of an instantiated HMM is presented in the survey [256]. Ghoshal et al. [78] have used a HMM for annotating images, by positing that image feature vectors describing low-level image content may be stochastically generated by a HMM, the states represent the keywords of interest. Wang et al. [301] pointed that human beings tend to view images as a whole. Thus, some semantic concepts cannot be learnable through single regions. The relationships between regions have also been considered for the semantic indexing of images using 2-D HMM for image annotation. Senthilkumar et al. [247] introduced a method to annotate images with keywords from a generic vocabulary of concepts or objects for the purpose of content-based image retrieval. The suggested method is based on HMMs for an automatic annotation and annotation-based image retrieval. In the automatic annotation task, a *Semantic annotated Markovian Semantic Indexing* (SMSI) is introduced. It consists in modeling the images, represented as sequence of feature vectors characterizing low-level visual features, like color, texture and oriented-edges, as having been stochastically provided by a HMM, whose states represent concepts. The parameters of the model are estimated from manually annotated (training) images. Then, each image from a large test collection is automatically annotated with a posteriori probability of concepts present within it. Image Annotation Using Spatial HMM is a 2-D generalization of the traditional HMM in the sense that both horizontal and vertical transitions between hidden states are taken into consideration. After annotating images, semantic retrieval of images can be performed by using Natural Language processing tool, namely *WordNet*, and measuring semantic similarity of annotated images in the database by using *Markovian Semantic Indexing* (MSI) [22].

- The *Artificial Neural Network* (ANN) is an unsupervised machine-learning algorithm inspired from the biological way of information processing of the human brain. It is able to learn from examples and provide decisions about new samples. ANN is credited for its ability to learn multiple classes all at once. An ANN consists of three layers, namely: *input, hidden* and *output*. Each layer consists of nodes (or neurons) performing numerical computations and other operations. Each neuron from a layer is interconnected with other neurons presented in consecutive layers. There are a bias assigned to each layer and a weight assigned to each interconnection. Fig. 13 shows a simple neural network. The input layer has neurons equal to the dimension of input sample. It is responsible for receiving large volumes of data as inputs in different formats (text, images, csv files, etc.). The output layer is responsible for producing the target outputs. All the calculation are performed in the hidden layer. Indeed, each neuron from the hidden layer operates as a processing element. It is governed by an activation function, which provides output according to the weights of the connecting edges and the outputs of the neurons of the previous layer. During the training process, a NN learns the edge weights in order to minimize the overall learning error. To classify a new sample, each output neuron generates a confidence measure. The class that corresponds to the maximum measure indicates the decision about the sample. Fig. 14 shows the procedure of the simple Neural Network. Hambali et al. [93] proposed a fruit classifier using a simple neural network model. The main aim of this study is to categorize '*jatropha fruits*' according to their color features. The input layer consists of six neurons $\{x_1, x_2, \ldots, x_6\}$, where each neuron represents the color of the elements ; R, G, B, L*, A* and B*, respectively. The input layer receives the signal, and then distributes the signal to the neurons in the hidden layer. The number of neurons in the hidden layer is seven and it is assumed sufficient to generate good prediction results. The output layer consists of four neurons, $\{t_1, t_2, t_3, t_4\}$, which represent the quality that a

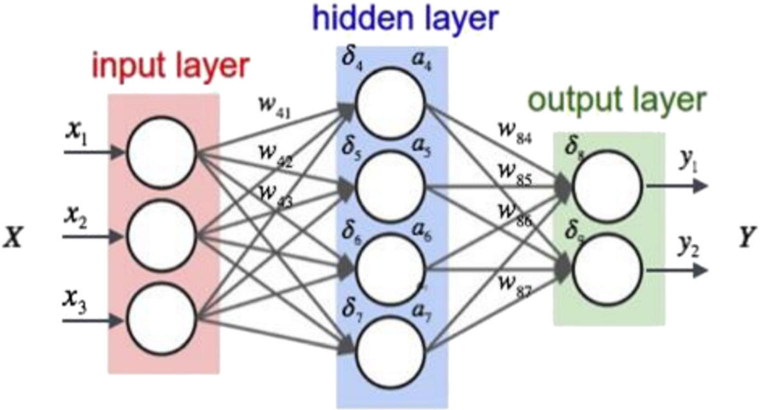

Fig. 13 A simple Neural Network

fruit can have ; '*immature*', '*under mature*', '*mature*' and 'o*ver mature*', respectively. Park et al. [216] suggested a method of content-based image classification using a 3-layer ANN, where the hidden layer consists of 49 neurons. The images for classification are object images, which can be divided into background and foreground. Thus, a pre-processing step is proposed for segmenting an image into a set of regions. The largest region at the centre of the image is used to identify the image. The regions with similar color distribution to the central region are considered as foreground (objects) regions. The foreground regions are used in order to extract the statistical texture features, which are transmitted to the ANN to classify the image into one of 30 concepts. In the study of Kaya et al. [138], the Butterfly dataset, containing 140 butterfly images, is divided into

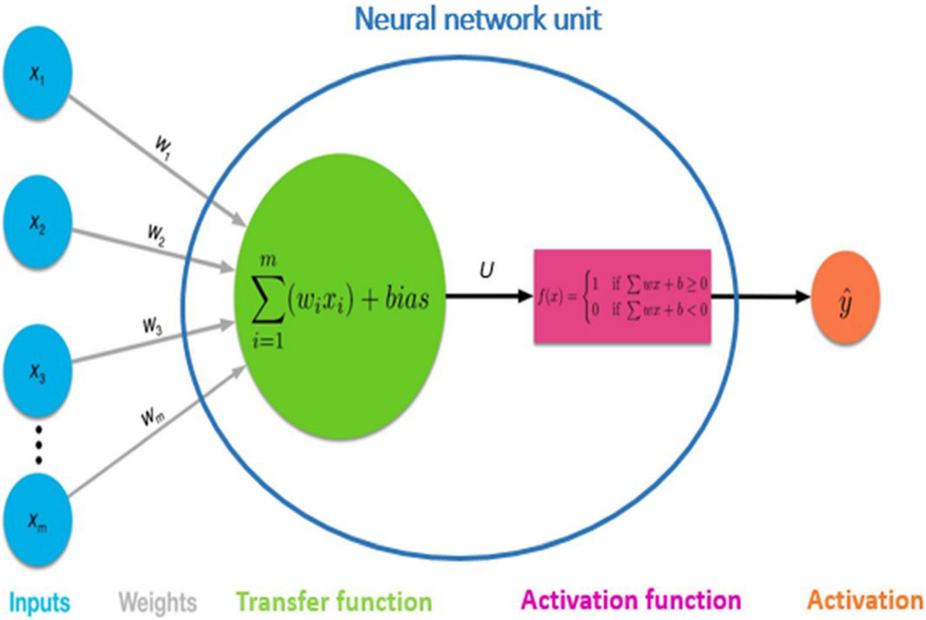
Neural network unit
x_1
x_2
x_3
x_m
w_1
w_2
w_3
w_m
$\sum_{i=1}^{m}(w_i x_i) + bias$
U
$f(x) = \begin{cases} 1 & \text{if } \sum wx + b \geq 0 \\ 0 & \text{if } \sum wx + b < 0 \end{cases}$
$\hat{y}$
Inputs
Weights
Transfer function
Activation function
Activation

Fig. 14 Procedure of the simple Neural Network

14 classes of different butterfly species of Styridae family. A pre-processing and tessellation step is implemented for resizing the image to (256×256) pixels. Five texture (correlation, contrast, entropy, homogeneity and energy) and three color (average of R, G, and B color bands) features of images are extracted. Each texture feature is calculated for different GLCMs (orientations are 0°, 45°, 90°, and 135°) and the distances (d = 1, 2, 3, 4) parameters that are established from the butterfly images. The average value of the texture features are calculated and used with the color features as input for the nodes of the ANN. Therefore, the number of input neurons is equal to the number of features of the data set. The number of hidden layers is 1, that of hidden neurons is 50 and that of output neurons is 14. In their study, Kuroda et al. [150] used four different 3-layer ANNs in order to hierarchically classify image regions. The numbers of neurons used in the hidden layers of these networks are 30, 10, 20, and 20, respectively. In this classifier, an image is first composed into some regions, and then each region is roughly classified into three broad categories, namely : '*sky*', '*water*', and '*earth*', by using SEW neural network. Second, the image features are extracted from each of the category, and the impression words (like '*bright*/*dark*', '*heavy*/*light*', '*warm*/*cool*', '*emotional*/*reasonabl*e' and '*rura*l/*urban*') are estimated from the image by using the second neural network called IW network. The regions belonging to sky or earth categories are classified into much more detailed objects, such as '*blue sky*', '*cloud*', '*sunset*', '*mountain*', '*green*' and '*rock*', by using the OR neural network. The fourth neural network does not classify any region, but it permits to associate an image with a vector of 18 dimensions and each dimension measures the degree of certain global characteristics of the image, like '*bright*/*dark*', '*rural*/*urban*' and '*busy*/*plain*'.

4.3.2 Deep learning

The remarkable progress of the hardware technologies as well as the explosive growth and availability of data have guided to the emergence of new machine-learning technique called *Deep-Learning* (DL). DL has its roots from artificial neural networks and significantly outperforms its predecessors. It uses graph technologies with transformations between neurons to develop learning models that consist of many hidden layers, as shown in Fig. 15, where the name *Multi-Layer Perceptron* (MLP). The choice of the number of hidden layers as well as the number of neurons at each hidden layer are open issues in the DL approaches.

Traditionally, the efficiency of machine-learning algorithms has been highly related to the quality of input data representation. Indeed, a bad data representation often degrades the quality of the results produced by a *machine-learning* and leads to lower performance compared to a good data representation. Thus, feature engineering has been for a long time considered as an important research direction in ML. In fact, it focuses on building features from raw data, which expanded the research studies. In addition, feature engineering is usually a very specific domain that requires significant human efforts. Once a new feature is suggested and proven effective, it will be a trend for years.

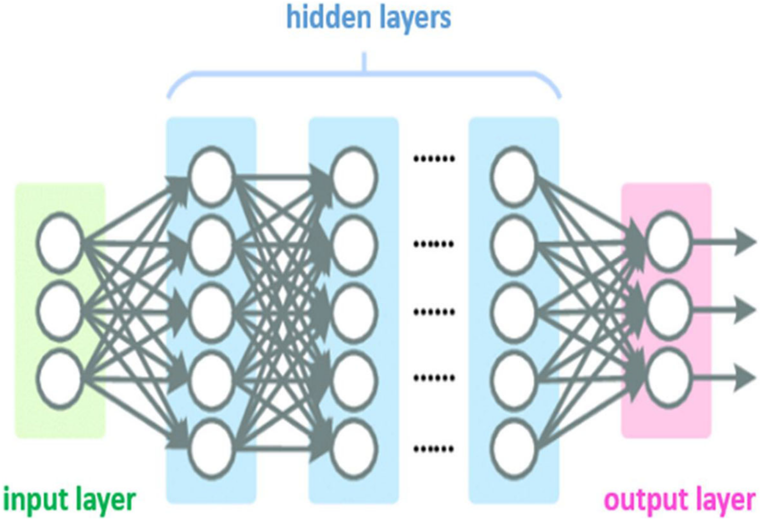

Fig. 15 A Deep-learning Neural Network

Deep-Learning algorithms offer the possibility to perform feature extraction with automatically way. It helps researchers extract discriminative features even with minimal domain information and reduce the human efforts [204]. These algorithms are characterized by a layered architecture of data representation where the low-level features are extracted from the lower layers and the high-level features are extracted from the last layers of the networks. This type of architecture has been originally inspired from the *Artificial Intelligence* (AI), which simulates its process of the key sensorial areas of the human brain. Indeed, a human brain can automatically extract data representation from different scenes. The inputs are the scene information received from eyes and the outputs are the objects classified by the brain. This highlights the major advantage of *deep-learning* that mimics how the human brain operates. Fig. 16 shows the feature extraction stage in traditional *machine-learning* and *deep-learning*.

Many DL techniques have been suggested in the literature. They have demonstrated promising results through different categories of applications, such as: *Recursive Neural Network* (RvNN) for Natural Language Processing (NLP) [262], *Recurrent Neural Network* (RNN) for NLP [102] and speech processing [245], and *Deep Boltzmann Machine* (DBM) for speech processing [77] and object recognition task [171]. *Convolutional Neural Network* (CNN) is also a popular and extensively used algorithm in DL. It has been widely applied into different applications, such as speech processing [1], NLP [304] and computer vision, spatially object recognition [80, 101], image annotation and retrieval [193, 306] and image captioning [9, 88, 300]. Similar to the traditional ANNs, the structure of the CNN is inspired from the neurons in human and animal brains. More specifically, it simulates the visual cortex in a biological brain containing arrangements of simple and complex cells. As shown in Fig. 17, these cells are sensitive to sub-regions of the visual field rather than to the whole scene. These sub-regions are named receptive fields.

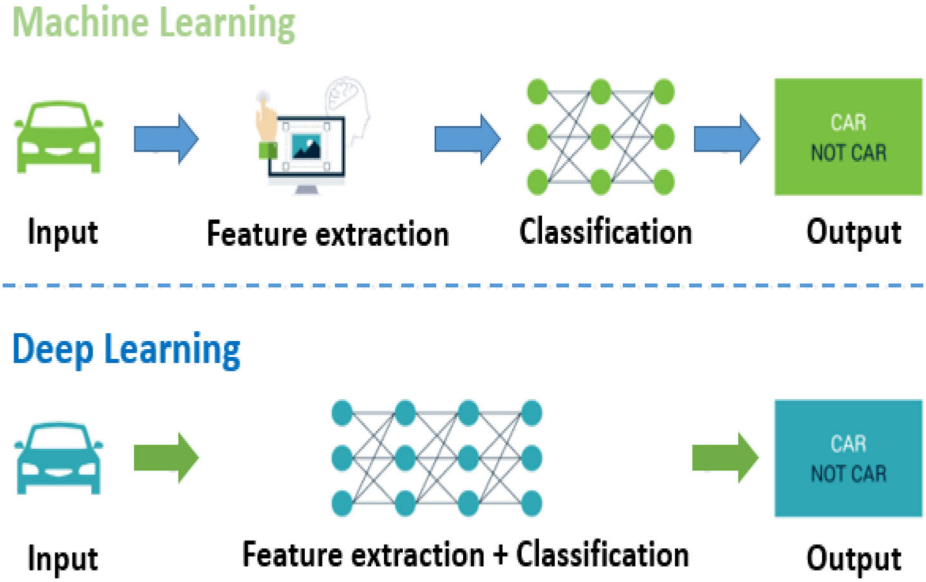

Fig. 16 Comparison between traditional machine-learning and deep-learning techniques

Similarly, neurons in a convolutional layer of a CNN connect to the sub-regions of the layers before that layer rather than being fully-connected like in other models of NNs. The neurons do not respond to the zones situated outside of these sub-regions in the image. These sub-regions might overlap so the neurons of a CNN generate spatially-correlated outcomes. However, the neurons do not share any connections and provide independent outcomes in other types of NNs. In addition, in NNs with fully-connected neurons, the number of parameters (or weights) can increase as the size of the input increases. A CNN

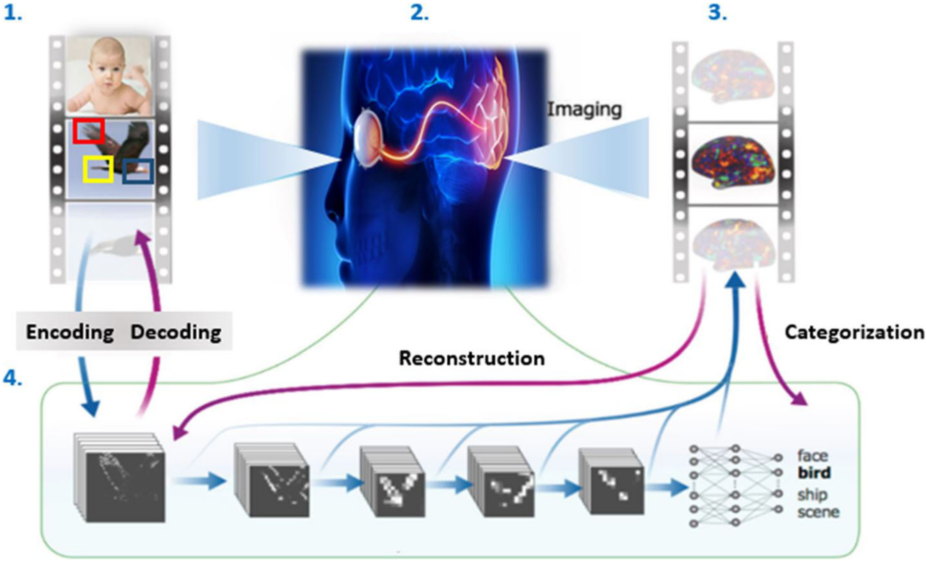

Fig. 17 Neural network inspired from the human brain

helps reduce the number of parameters by minimizing the number of connections, shared weights, and down-sampling.

A typical CNN consists of a number of convolutional layers followed by pooling layers (sub-sampling), which are responsible for the feature extraction. The learned features become inputs to the fully-connected layers for the classification task. Fig. 18 shows the typical structure of a CNN.

In each layer of a CNN, the neurons are arranged in a 3-D manner enabling to transform a 3-D input to a 3-D output. More specifically, as shown in Fig. 19, the layers in a CNN have the inputs *x* arranged into three dimensions (*m*×*m*×*r)*, where *m* refers to the width and height of the input, and *r* refers to its depth or channel numbers (e.g., *r* = 3 for RGB images).

In the convolutional layers, there are several filters (kernels) *k* of size (*n*×*n*×*q*). Indeed, *n* can be smaller than the original image, but *q* should be smaller or the same size as *r*. The kernels *k* are the base of local connections that are convolved with the input. They share the same parameters (bias b^k and weight W^k) to produce *k* feature maps (h^k). Similar to the MLP, the convolutional layers compute the dot product between the weights and its inputs, as illustrated in Equation (2), only the inputs are small regions of the initial input volume.

$$h^k = f\left(W^k * x + b^k\right) \tag{2}$$

In the pooling layers, each feature map is down-sampled to reduce the parameters in the network, accelerate the training process, and therefore control the overfitting. The pooling operation (e.g., max or average) is performed on a (*p*×*p*) contiguous region for all feature maps, where *p* is the filter size.

The layers of the fully-connected stage receive the resulting low-level and mid-level features as inputs and provide high-level abstractions for the processed image. The last layer (e.g., Softmax or SVM) should be used for generating classification scores. Each score represents the probability of some class for a given instance.

The general CNN model implementation can be presented as shown in Fig. 20.

CNN model helps perform a hierarchical feature representation, which can be automatically learned from data. Compared with the traditional descriptors, CNN offers a deeper architecture that can provides an exponentially evolved expressive capability. This architecture also

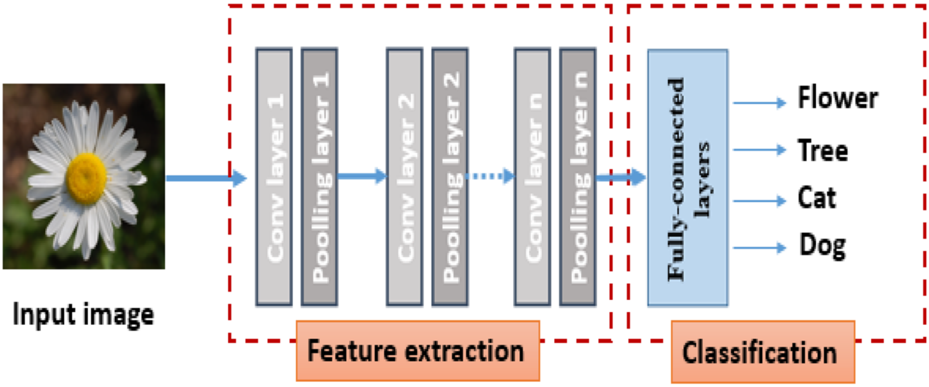

Fig. 18 Structure of a typical CNN

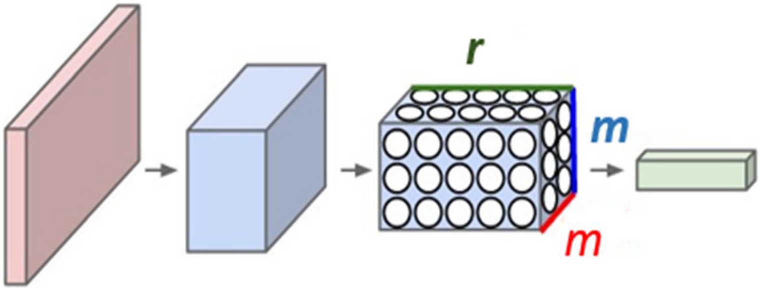

Fig. 19 A layer of CNN in 3 dimensions

provides the opportunity to jointly optimize numerous related tasks together, such as bounding box regression, feature extraction and classification. Thanks to its interest, CNN model has been frequently applied into many research fields, such as image classification [146], image annotation and retrieval [193, 306], object recognition [80, 101], image captioning [9, 88, 300], etc. For example, in Ref. [59], the deep CNN model that won the ImageNet Challenge in 2012, was employed for large-scale image classification and object recognition. The suggested neural network architecture consists of 650,000 neurons with 60 million parameters containing five convolutional layers. Some of these layers have been followed by max-pooling layers and three fully-connected layers including a final 1000-way Softmax layer. Combined with new techniques, such as *Dropout*, *Rectified Linear Units* (ReLUs) and a very efficient GPU implementation, the suggested CNN model has achieved in the ILSVRC-2012 competition a winning top-5 test error rate of 15.3%. However, it is concerned with single-label image classification.

In order to generate multi-label image annotation, Gong et al. [84] have suggested to use the same CNN model proposed by Deng et al. [59] and mainly focus on training the network with loss functions adopted for multi-label prediction tasks. The first step of the proposed annotation process consists in resizing each image to (256×256) to be sent to the convolutional layers. Thereafter, (220×220) patches are extracted from the entire image, one from the center and four from the corners, to generate an augmentation of the dataset. Convolution kernel sizes are respectively set to squares of size 11, 9, and 5 for the different convolutional layers. Max-pooling layers are employed in some of the convolutional layers to offer invariance. Each fully-connected layer has output sizes of 4096 and it is followed by *Dropout* layers. For all the layers, ReLU is used as non-linear activation function. For minimizing the multi-label Softmax regression loss, a first loss function, inspired by Tagprop [90], has been used. The second used loss is a simple modification of a pairwise-ranking loss [128] enabling to take multiple labels into account. The third loss function consists of a multi-label variant of the WARP loss [311]

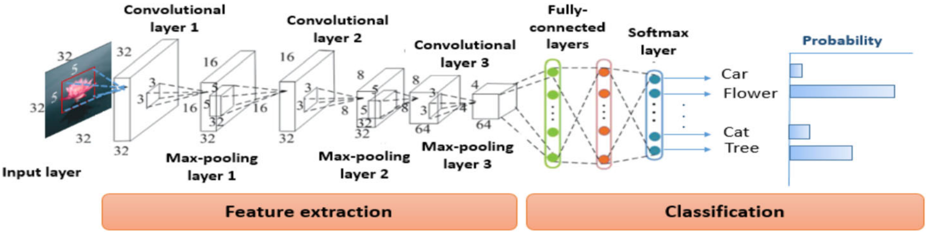

Fig. 20 Example of a CNN model implementation

enabling to use a sampling trick in order to optimize top-k annotation accuracy. For comparison purposes, nine different visual features, including, GIST, D-SIFT, H-SIFT, HOG, have been used as baseline features. Two powerful classifiers, namely: SVM and *k*-NN, have been implemented for annotating image according to these features. The experimental results demonstrated that the deep CNN outperforms the existing visual-feature-based methods in image annotation. However, this type of model needs to be recycled when new labels are released.

In the study of Mayhew et al. [193], two image annotation algorithms, called TagProp [90] and 2PKNN [294], were trained with features derived from the two CNN architectures (VGG-16 and AlexNet). Experimental results have proved that the annotation performance reached by using features derived from a deep CNN outperforms the one that is based on larger handcrafted features.

In Ref. [48], multi-label image annotation using deep CNN model was also suggested by Chengjian et al. The overall architecture of the implemented CNN model consists of eight layers with weights. The first five represent the convolutional layers and the remaining three are the fully-connected layers. The outputs of the fully-connected layers are introduced into a 1000-way Softmax classifier that produces a distribution over 1000 labels. In implementation stage, the global features that were extracted as visual descriptors are: (1) 128-dimension HSV color histogram and 225-dimension LAB color moments; (2) 37-dimension edge direction histogram; (3) 36-dimension Pyramid Wavelet texture; (4) 59-dimension local binary pattern feature descriptor, and (5) 960-dimension GIST feature descriptor. The local features that were extracted as visual descriptors are: (1) a Harris corner detector and the dense sampling method that were adopted as patch-sampling methods; (2) SIFT feature, CSIFT feature, and RGBSIFT feature that were extracted to form a codebook of size 1000 using *k*-means clustering; (3) a two-level spatial pyramid that was adopted to construct a 5000-dimensional vector for each image; (4) the TF-IDF weighing scheme that was utilized to generate the final bag-of-visual-words. For the all experiments, the feature vectors were normalized to the range of [0, 1]. Experimental comparisons were performed between the proposed image classification method and: (1) Lazy learning based approach (LL) [333]; and (2) Deep representations and codes based approach (DRC) [142]. The results of the evaluations have demonstrated that the proposed deep-structured semantic model considerably outperforms the two other approaches for three used image datasets.

Murthy et al. [202] proposed a CCA-KNN model based on the *Canonical Correlation Analysis* (CCA) framework. The new framework helps model both textual features (word embedding vectors) and visual features (CNN features) of the data. The CNN features have been proven more efficient compared with 15 handcrafted features from existing models, which include JEC [182], SVM-DMBRM [201], TagProp [90] and 2PKNN [294]. Moreover, this study has shown that the word embedding vectors performed better than the binary vectors as a representation of tags associated with an image.

Wang et al. [306] have suggested *a Multi-task Voting Automatic Image Annotation CNN* model (MVAIACNN) that helps interrogate training and test datasets. The *Multi-task Voting* method (MV) helps achieve the adaptive label by combining the multi-task learning method with the *Bayesian* probability model. Thereafter, a (AIACNN) model have been proposed. It consists of five convolutional-layers for hierarchically extracting features, and four pooling-layers followed by two fully-connected layers and a Softmax output layer that defines identity classes.

Wu et al. [315] designed a framework, called *Deep Multiple Instance Learning* (DMIL) model, which helps learn the coincidences between image regions and keywords. In DMIL framework, object proposals and keywords are simultaneously learned based on a joint deep multi-instance learning. Indeed, DMIL uses a CNN model that consists of five convolutional layers, a pooling layer and three fully-connected layers. Thereafter, another deep neural network framework, which contains one input layer, one hidden layer, and one output layer with a Softmax for multi-instance learning is used. Finally, the text outputs and image in the fully-connected layer are combined.

4.3.3 Summary (comparison between ML and DL algorithms)

Table 4 presents the advantages and disadvantages of the various ML and DL algorithms that have been studied.

4.3.4 CNN-based object recognition frameworks

In essence, object recognition aims at detecting and classifying the objects depicted on any one image, and labeling them with rectangular bounding boxes (anchor boxes), as shows in Fig. 21. Indeed, the labels attributed to the detected objects contributed for the image annotation process. The pipeline of object recognition is divided into three steps:

- *Object proposal*: The main purpose of this step is to search within a given image the locations that can contain objects by scanning the whole image based on sliding windows [56, 97, 291]. To detect information about multi-scale objects, input images are resized into different scales. Multi-scale windows are also employed to slide via these images.
- *Feature vector extraction*: For each location detected on the image, a fixed-length feature vector is provided from the sliding windows in order to get discriminative semantic information about the covered region. The feature vector is usually encoded by using a CNN model.
- *Region classification*: Categorical labels are associated with the covered regions by using region classifiers. *Support Vector Machines* (SVM) are the most commonly used classifiers.

The frameworks of the generic object recognition methods based on the CNN model can mainly be categorized into two types: *region proposal-based frameworks* and *regression/classification-based frameworks.*

1) The *region proposal-based frameworks* follow the classical object detection pipeline, which first generates region proposals and thereafter classify each proposal into different object types. Among the most popular *region proposal-based frameworks* are R-CNN [80], SPP-net [100], Fast R-CNN [79], Faster R-CNN [236], R-FCN [54] and Mask R-CNN [101]. The contract of the different *region proposal-based frameworks* is introduced in Table 5.

- **Regions with CNN features (R-CNN):** Girshick et al. [80] proposed a novel CNN architecture that helps improve the quality of the candidate bounding boxes and extract the top-level characteristics by using a deep architecture. The flowchart of the R-CNN architecture consists of three steps such that :

Table 4 Advantages and Disadvantages of the different ML and DL algorithms that have already been studied.

Algorithm	Advantages	Disadvantages
k-NN	1. Manipulate non-parametric training data, 2. Training step is very fast, 3. Simple to learn, 4. Robust to noisy training data, 5. Effective when training data is large.	1. Biased by the value of k, 2. Computation are complex, 3. Limitation of the memory, 4. Testing step runs slowly.
DT	1. Manipulate non-parametric training data, 2. Does not required an extensive training 3. Generates the deep learning features hierarchical associations between input variables to predict class membership and produces a set of rules that are easy to interpret, 4. Simple and efficient computational.	1. The computation becomes complex when various outcomes are correlated and/or various values are undecided.
SVM	1. Achieves optimal class boundaries by finding the maximum distance between classes, 2. Provides a good generalization capability, 3. The adjustment problem is eliminated, 4. Computational complexity is reduced, 5. Simple to manage the error frequency and decision rule complexity.	1. Result transparency is weak, 2. Training step is time consuming, 3. Structure of the algorithm is difficult to understand, 4. Determination of optimal parameters is complex when there is non-linearly separable training data.
BN	1. Performance is good, 2. Easy to implement, 3. Takes less computational time for processing.	1. The dependencies existing between variables are ignored, which would cause it to provide less accurate predictions.
Partitioning clustering	1. Simple and relatively scalable, 2. Appropriate for datasets with compact spherical clusters, which are well-separated.	1. Serious effectiveness degradation in high dimensional spaces. 2. Poor description for clusters. 3. Requires a manual specification of the number of clusters in advance. 4. High sensitivity to initialization phase, outliers and noise. 5. Frequent entrapments in the local optima.
Hierarchical clustering	1. Embedded flexibility concerning the level of granularity, 2. Well adapted for problems that involve point linkages, such as taxonomy trees.	1. Inability to perform corrections once the splitting or merging decision is made, 2. Cloudiness of termination criterion, 3. Expensive for massive and high dimensional datasets, 4. Serious effectiveness degradation in case of high dimensional spaces.
HMM	1. Allows an efficient learning that can be performed directly from raw sequence data.	1. Not completely automatic and requires training using annotated data, 2. The size of training data can be an issue.
ANN	1. Enables to manipulate non-parametric training data, 2. Capability to present functions, such as AND, OR and NOT, 3. Consists of data driven self adaptive technique, 4. Efficiently handles noisy inputs, 5. Computation rate is important.	1. Semantic poverty, 2. Problem of over-fitting, 3. The training of ANN is time consuming, 4. Difficult to define the network architecture.
CNN	1. Treats large data, 2. Process complicated relationships, 3. Derives robust characteristics, 4. No manual choice is needed, 5. Multi-labeling of images.	1. Optimum is local, 2. Training stage cannot be controlled, 3. Needs large training images.

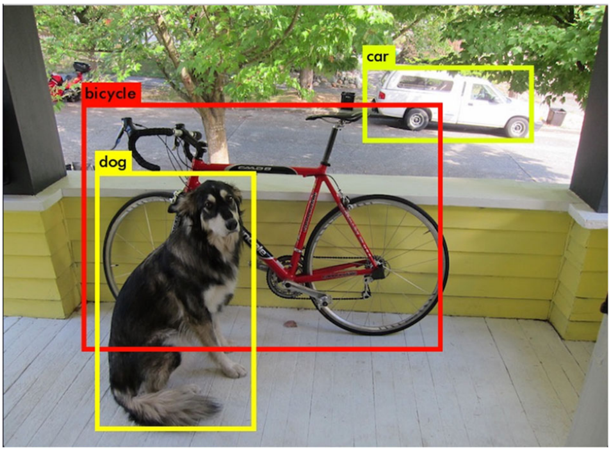

Fig. 21 Object recognition based on the CNN model (example)

- *Region proposal generation*: By adopting a selective search, the R-CNN model helps provide about 2k region proposals for each image. In fact, the selective search method [290] is based on simple bottom-up clustering and saliency indices enabling to quickly generate precise candidate boxes with arbitrary sizes for reducing the searching space of the object detection.
- *Deep feature extraction*: The generated region proposals are warped due to the fixed resolution. Thereafter, the CNN module [146] is applied to extract a 4096 dimensional feature as a final representation. The large learning capacity and the hierarchical structure of CNNs have enabled a high-level feature representation for region proposals.
- *Classification and localization*: Based on pre-formed SVMs classifiers, the different region proposals are labeled on a set of positive and negative regions. Thereafter, these regions are adjusted by using bounding box regression and filtered with a non-maximum suppression (NMS) in order to generate the final bounding boxes of the detected objects.

Despite the fact that the R-CNN model has achieved high accuracy rate versus the traditional methods, some drawbacks are noted. Indeed, due to the existence of Fully Connected layers (FCs), the CNN model requires a fixed-size input image (e.g., 256×256) to re-calculate the whole CNN for each evaluated region proposal, which is a time-consuming operation in the testing period. In addition, the R-CNN model distorts the generated region proposals and design them into the same size. However, unwanted geometric distortion may be produced, and consequently some object information can be lost. This distortion of content can reduce the recognition accuracy, in particular when the scales of objects vary. On the other hand, the generated region proposals are redundant although the selective search helps generate them with relatively high recalls, which is also a time-consuming procedure (2s to extract 2k region proposals). Besides, features are extracted from the region proposals and stored on the disk. Thus, a very long time may elapsed to access them and the storage memory required by these features can be also expensive. Finally, the R-CNN is multi-stage pipeline. Indeed, at the training of the R-CNN model, a convolutional network (ConvNet) on object proposals is finetuned. Thereafter, the softmax classifier learned by fine-tuning is replaced by an SVM one in order to fit in with (ConvNet) features. Finally, the bounding-box regressors are trained.

- **Spatial Pyramid Pooling in Deep Convolutional Networks (SPP-net):** To resolve the issue of the information distortion of the region proposals caused by the R-CNN model, He et al. [100] took into account the theory of the Spatial Pyramid Matching (SPM) [156] and [223], and proposed a novel CNN model called SPP-net. Indeed, the SPM takes many thinner to coarser scales in order to divide the image into a some number of partitions and regroup the quantized local characteristics into intermediate level representations. In addition, the R-CNN model is heavy because it realizes a ConvNet forward pass for each object proposal without dividing computation. Therefore, the SPP-net model proposed to speed up R-CNN by sharing computation. It computes a conv feature map for the whole input image. Thereafter, it classifies the object proposal set by using a feature vector extracted from the shared feature map. In fact, the features are extracted from a given region proposal by max-pooling the portion of the feature map inside the proposal into a fixed-size output (e.g., 8 × 8). The multiple output sizes are grouped and then concatenated as a spatial pyramid pooling. SPP-net speed up the R-CNN model by 10 to 100× at test time and reduce the training time by 3× thanks to the fast proposal feature extraction.

Despite its advantageous interventions, SPP-net shares some limitations with R-CNN, such as the multi-stage pipeline, including feature extraction, SVM training, network fine-tuning and bounding box regressor fitting. Therefore, the storage memory required is also expensive. Besides, the conv layers foregoing the SPP layer cannot be updated based on the fine-tuning algorithm. Thus, a decrease accuracy of deep networks has been noted.

- **Fast R-CNN:** Girshick [79] proposed a novel CNN architecture called Fast R-CNN to resolve the multi-stage pipeline issue by introducing a multi-task loss at the classification and bounding box regression steps. The R-CNN model takes as input a given image and the region proposal set. At first, it processes the whole image with conv and max pooling layers to generate a conv feature map. Thereafter, a region of interest (RoI) pooling layer extracts a fixed-length feature vector from the feature map for each object proposal. Each feature vector is then introduced into a sequence of FC layers, which finally branched into two output layers where the first one generates softmax probability that estimates over C object classes and a '*background*' one. The other output layer encodes the refined bounding-box positions based on four real-valued numbers. The parameters in these processes are optimized through a multi-task loss in an end-to-end way, exceptionally the generation of the region proposals.

The Fast R-CNN model also proposed an efficient training method, which takes advantage of feature sharing at the training step. In this context, we recall that if the region of interests come from different images, the back-propagation via the SPP layer becomes very inefficient. Therefore, Stochastic Gradient Descent (SGD) mini-batches are hierarchically sampled by sampling first X images and then Y/X RoIs from each image (Y is the number of RoIs). In the other words, the RoIs of the same image share memory and computation in the forward and backward passes. Besides, the truncated

Singular Value Decomposition (SVD) [319] is used in order to compress large FC layers and speed up the testing procedure.

Although the Fast R-CNN model permits to execute the training of the all network layers in a one-stage with a multi-task loss, reduce the additional expenses on storage memory, and improve both accuracy and efficiency thanks to robust training schemes, the time spent on region proposals is still ignored.

- **Faster R-CNN:** R-CNN and Fast R-CNN use a selective search to find out the region proposals. However, the selective search process is a slow and time-consuming process, which affects the performance of the network. Therefore, Ren et al. [236] defined an object detection algorithm that lets the network learn the region proposals and hence eliminates the selective search algorithm.

Similar to the Fast R-CNN model, an image is provided as an input to a convolutional network, which generates a convolutional feature map. However, instead of using the selective search algorithm on the feature map to identify the region proposals, Faster R-CNN model proposed to use a novel network named Region Proposal Network (RPN) in order to predict the region proposals. Thereafter, the predicted region proposals are reshaped by using a RoI pooling layer, which is also used to classify the image within the proposed region and predict the offset values for the bounding boxes. The Faster R-CNN model is much faster than its predecessors. Then, it can be used for real-time object detection. However, RPN produces object-type regions, including backgrounds, instead of object instances. Therefore, Faster R-CNN is not efficient in proceeding with very small objects.

- **Region-based Fully Convolutional Networks (R-FCN):** The RoI pooling layer of the Faster R-CNN detector is unnaturally inserted between two sets of convolutional layers due to the dilemma of increasing translation invariance for image classification versus respecting translation variance for object detection. Indeed, the framing of an object depicted on an image is random at the classification process while any translation of an object in a bounding box may be meaningful in object detection. Thus, fully convolutional architectures, that are translation invariant, are preferable. Besides, the object detection step requires localization representations, which are translation-variant to an extent. As instance, the translation of an object within a candidate bounding box must produce meaningful responses to describe how good the candidate bounding box overlaps the object. To address this dilemma, the RoI pooling layer was inserted into convolutional layers. However, this design can affect the training and testing efficiency because it introduces a considerable number of region-wise layers. It is also a costly per-region subnetwork hundreds of times.

In contrast to the previous region-based detectors, namely Fast/Faster R-CNN, Dai et al. [54] proposed position-sensitive score maps to address the dilemma between translation-invariance in image classification and translation-variance in object detection. With the Region-based Fully Convolutional Networks R-FCN, more effective classification networks can be used to perform the object detection procedure in a fully-convolutional architecture by sharing all the layers. A test speed of 170ms per image is achieved on both Microsoft COCO and PASCAL VOC datasets.

Table 5 Contrast of region proposal based frameworks

Frameworks	Pros	Cons
R-CNN	1. Selective search for region of interests, 2. Extracting high-level image features thanks to the use of a deep architecture.	1. Warping can loss object information, 2. Region proposal generation step is time-consuming, 3. Training is time consuming and expensive in terms of space, 4. Multi-stage pipeline, 5. Not suitable for real-time usage.
SPP-net	1. More significant in object proposals in their corresponding scales, 2. Speed up R-CNN by sharing computation, 3. Reducing the training time.	High memory consumption, Training is inefficient, Multi-stage pipeline.
Fast R-CNN	1. One fine-tuning stage, 2. Fast and efficient training, No disk storage is required for feature caching.	1. Ignoring the time spent on region proposals.
Faster R-CNN	1. Real-time object detection.	1. Limited success for detecting small objects, 2. Training is inefficient, 3. A costly per-region subnetwork.
R-FCN	1. Boosting classification and object detection, 2. Much faster during training and inference, 3. A robust and efficient feature extractor.	1. Requires a deep model over-fitting for the most real-word applications.
Mask R-CNN	1. Efficient solution for overlapping instances, 2. Good inference speed, 3. Good accuracy, 4. Intuitive and easy to implement, 5. Extension capability.	1. False alerts, 2. Missing labels.

- **Mask R-CNN:** Instance segmentation is challenging because it requires detecting all objects within an image and segmenting each instance. However, the execution of these two tasks independently can provoke a problem of overlapping instances. To solve this issue, He et al. [101] proposed to extend Faster R-CNN by adding a novel branch that can predict segmentation masks in a pixel-to-pixel way. More specifically, Mask R-CNN predicts an m × m mask from each RoI by using an FCN [175]. This helps each layer in the mask branch to maintain the explicit m × m object spatial layout without collapsing it into a vector of representation that miss spatial dimensions. As demonstrate the performed experiments, the proposed FC representation for mask prediction requires fewer parameters but it is very accurate. However, the proposed pixel-to-pixel approach requires the aligned of the RoI features in order to preserve the explicit per-pixel spatial correspondence. This motivated authors to develop a RoIAlign layer.

Indeed, the RoIAlign consists in avoiding any quantization of the RoI boundaries. Thus, a bilinear interpolation [116] was used for computing the exact values of the input features at four regularly sampled locations in each RoI bin and then aggregate the result. This has led to large improvements.

In essence, the Mask R-CNN model is an extension of Faster R-CNN where the construction of the mask branch helps rapidly achieving good results in terms of object detection by cooperating with other tasks, and adding only small computational burden allowing a fast system. The Mask R-

CNN is also simple to implement and train given the Faster R-CNN model that facilitates a wide range of flexible architecture designs. However, the Mask R-CNN has the disadvantage of sometimes generating false alerts and some labels may also be missing.

2) The *regression/classification-based frameworks* consists of many correlated steps, which are often trained separately. These steps include region proposal generation, feature extraction with CNN, classification and bounding box regression. The most popular *regression/classification-based frameworks* are MultiBox [64], YOLO [235], SSD [172], YOLOv2 [234], DSSD [72] and DSOD [252]. The contrast of the different *regression/classification-based frameworks* is introduced in Table 6.

- **MultiBox:** Erhan et al. [64] proposed a regression based MultiBox. The model helps achieve a class-agnostic scalable object detection by predicting a set of bounding boxes that represent potential objects. More precisely, a Deep Neural Network (DNN) is used in order to produce a fixed number of bounding boxes. Besides, a score is attributed to each box containing an object in order to express the network confidence.

Although the proposed algorithm manages to produce large number of objects, an additional boost when using higher resolution image crops has been noted. In addition, the achieved Average Precision (AP) was not too satisfactory.

- **You only look once (YOLO):** In essence, YOLO [235] consists in dividing each image into a grid of S x S where each from them predicts N bounding boxes and confidence. The confidence score reflects the accuracy of the bounding box and whether the bounding box really contains an object regardless of class. YOLO also helps predict the classification score for each box for every class in training. The classes can be combined in order to calculate the probability of each class being present in a predicted box. Thus, a total of S x S x N boxes are predicted.

YOLO consists of 24 conv layers and 2 FC layers where some conv layers build sets of inception modules with 1 × 1 reduction layers followed by 3 × 3 conv layers. It can treat images in real-time. Furthermore, it can cooperate with Fast R-CNN because it provides fewer false positives on background. However, YOLO can generate localization errors of bounding boxes. It has relatively low recall compared to region proposal-based methods. In addition, YOLO has a difficulty in processing small objects in groups due to strong spatial constraints imposed on bounding box predictions.

- **Single Shot Multibox Detector (SSD)**: Liu et al. [172] introduced the SSD model, which is a single-shot detector for multiple categories. The SSD is a simple model compared to methods that require object proposals because it completely eliminates proposal generation and subsequent pixel or feature resampling stages and encapsulates all computation in a single network. Indeed, the core of SSD helps predict category scores and box offsets for a fixed ensemble of default bounding boxes by using small convolutional filters applied to feature maps. To achieve high detection accuracy, predictions of different scales from feature maps of different scales are generated and explicitly separated by aspect ratio. Even on low resolution input images, these design features lead to simple end-to-end training and high accuracy, further improving the speed vs accuracy trade-off. It should be mentioned that the SSD model is faster than the previous state-of-the-art for single shot detectors (YOLO) and significantly more accurate than the slower techniques that perform explicit region proposals and pooling (including Faster R-CNN). However, the main drawback of SSD that is not able to deal with small objects.

- **YOLOv2:** Proposed by Redmon et al. [234], YOLOv2 constitutes the second version of the YOLO with the objective of improving significantly the accuracy while making it faster. In fact, by adding batch normalization on all of the convolutional layers in YOLO, more than 2% improvement in mAP is achieved. In addition, a high resolution classification network was proposed. This has given an increase of almost 4% mAP. YOLOv2 also tries to use the idea of anchor boxes by finding the best anchor boxes shapes to make it easier for the network to learn detection.

To make YOLOv2 robust to running on images of different sizes, the model was trained for different input sizes. Since the model adopts only convolutional and pooling layers, the input can be resized on the fly. Indeed, instead of fixing the input image size, authors proposed to change the network every few iterations. After every 10 batches, the network chooses a new image dimension size from a dimension set, randomly. Then, the network is resized to that dimension and continue training. Thus, the same network can predict detections at different resolutions (input shapes).

Despite all of its advantages versus the state-of-the art networks, such as Faster R-CNN and YOLOv1, YOLOv2 has a very complex architecture. In addition, YOLOv2 is not skilled at dealing with small objects.

- **Deconvolutional Single Shot Detector (DSSD):** In the original SSD architecture, preset anchor boxes are used to replace regional proposal generation. In addition, SSD uses more than one feature map for the detection in order to account for different object sizes. Despite its astonishing performance, SSD suffers from unsuccessful detection for small objects, which are typically defined as less than 36x36 within an image and then tremendously reduced being passed via multiple pooling layers. Therefore, the detection model in SSD does not have enough spatial information in order to discern the small objects.

To achieve greater performance, Fu et al. [72] proposed to improve both feature extraction and detection parts of the SSD model. At the first, they proposed to use the ResNet-101 feature extractor instead of the VGGNet-19 extractor used in the original paper. Indeed, empirical evidences have shown that the residual networks are easier to optimize and can gain accuracy from considerably increased depth.

Given that in the SSD network the feature maps are lightly processed before a loss function is applied to it, the feature extractor should learn to provide feature maps that not only represent the semantic and spatial information from the precedent layers, but also the transformation set that leads to good classification. However, different branches in SSD correspond to different scales. Therefore, it can be necessary to avoid previous transformations before applying the one that works best for its scale. In this regard, prediction modules (PMs) were used in the DSSD architecture to perform the necessary processing of feature maps and achieve a good classification. Experimental studies have demonstrated that (PMs) have increased the performance of the DSSD network.

Furthermore, deconvolutional layers are used in the detection part in order to increase the resolution of the feature maps generated by the feature extractor. Thus, a better detection of small objects by providing additional large-scale context has been achieved.

Although the DSSD has led to significant results in detecting small objects, the architecture of the model is complex and its speed is slow.

- **Deeply Supervised Object Detector (DSOD):** State-of-the-art object recognition architectures rely deeply on the off-the-shelf networks pre-trained on large-scale classification databases, such as ImageNet. Because of the difference on both the loss functions and the category distributions between classification and detection tasks, learning bias is induced. A model fine-tuning for the task of detection can mitigated this bias to some extent but not so much. In addition, transferring pre-trained models from classification to detection between divergent domains is even more complicated. To tackle these two critical issues, a best solution consists in training object detectors from scratch.

In this context, Shen et al. [252] proposed the DSOD framework that can learn object detectors from scratch. Indeed, previous efforts in this direction mainly failed because of the complexity of the loss functions and the limited training data in object detection. In DSOD, the authors contributed a set of design policies for training object detectors from scratch. It uses implicit Deep Supervision (DS), namely DenseNet, in order to mitigate vanishing gradients. Indeed, the idea of DS consists in bringing the loss function that is generally attached to the top part of the network, closer to different layers of the said network. This helps each layer to adopt a less diluted gradient to learn. Indeed, DenseNet consists of four dense blocks that maintain the same scale of outputs. To increase network depth, a solution consists in adding layers inside each block for the original DenseNet. The transition without pooling layer circumvents this restriction, enabling hence to use more dense blocks.

A stem block is also adopted to modify the original architecture of DenseNet. Indeed, in DSOD architecture a stack of 3x3 convolution layers followed by a 2x2 MaxPooling is used instead of using a 7x7 convolution layer with stride 2 followed by a 3x3 MaxPooling operation with stride 2. The first convolution layer has stride 2 while the others use stride 1. This can minimize information loss from the raw input image because the smaller filter sizes and strides tend to preserve information.

In order to improve the detection accuracy in DSOD, the feature maps processed from previous layers are concatenated with down-sampled, which are high-resolution feature maps in a one-to-one ratio for detection. In fact, processed feature maps from previous layers have information useful for classification while the high-resolution feature maps preserves spatial information. For comparing, the SSD adopts only feature maps processed from previous layers in order to make multi-scale detection.

DSOD is a simple and efficient framework for training object detector from scratch. It has great potential on different domains, such as medical images, etc. Its lightweight structure helps achieve a 77.7% mAP on the VOC 2007 test set without pre-training. However, DSOD does not promote the performance when using pre-trained network. Table 6

4.3.5 Image captioning

Image captioning consists in recognizing the important objects depicted in any one image, their attributes, and their relationships, as shows in Fig. 22. The different information that reflect the visual and semantic content of images contribute to their annotation. However, image captioning requires generating semantically and syntactically correct sentences. The different approaches of image captioning can be divided into three categories: *template-based image captioning* [67, 148, 163], *retrieval-based image captioning* [85, 104, 213, 268] and *deep-learning-based image captioning* [66, 126, 300, 318]. Indeed, image captioning challenges have been successfully processed by using deep-learning-based techniques.

Deep-learning-based image captioning methods can use *visual space* [43, 66] or *multimodal space* [143, 144] for mapping image features. They can also be categorized according to the learning techniques: *supervised learning* [135, 188], *reinforcement learning* [237, 238], and *unsupervised learning* [55, 253]. Captions can be generated for a *whole scene* [122, 134] or for different regions of an image (*dense captioning*) [130, 324]. Image captioning methods can be based on a *simple encoder-decoder architecture* [143, 144] or a *compositional architecture* [66, 293]. There are methods that adopt *attention mechanisms* [126, 318], *semantic concepts* [280, 328], *stylized* [73, 192] and *novel object captioning* [293, 325]. Some image captioning methods use CNN as a language model [9, 88, 300]. Nevertheless, other language models have been used, such as TSLM [180, 267], RNN [189, 341], LBL [143], DTR [135] and MELM [43, 283]. Fig. 23 illustrates the taxonomies of the different deep-learning-based image captioning methods.

For more details on the different deep-learning-based image captioning methods readers can refer to [108].

For image captioning purposes, Aneja et al. [9] proposed a convolutional architecture, which consists of four components: an input embedding layer, an image embedding layer, a convolutional module, and an output embedding layer. Spatial soft attention mechanism has been also employed. The experimental results have demonstrated that the proposed architecture provides comparable performance versus a LSTM-based method by using standard metrics on the challenging MSCOCO dataset. Wang et al. [300] proposed a CNN+CNN-based image captioning method, which consists of the similar architecture proposed by Aneja et al. [9]. To improve the performance of the image captioning method, authors use a hierarchical attention module for connecting the vision CNN with the language CNN. They also study the influence of the hyper-parameters, namely: the number of layers and the *kernel* width of the language CNN. The experimental results demonstrated that the hyper-parameters help improve the performance of the image captioning method. Gu et al. [88] proposed an image captioning method, where they combine the RNNs model with the CNN language to highlight the temporal dependencies.

4.3.6 Differences between machine-learning and deep learning

Deep-learning (DL) is a special type of *Machine-learning* (ML), which is also a subfield of *Artificial-intelligence* (AI).A subset representation of the learning algorithms is illustrated in Fig. 24.

Table 7 summary the main differences that exist between *machine learning* and *deep learning.*

Table 6 Contrast of regression/classification based frameworks

Frameworks	Pros	Cons
MultiBox	1. Capacity to capture multiple instances of objects of the same class, 2. Scalable, 3. Lower computational cost.	1. Supplementary parameters are introduced to the final layer, 2. Not very good accuracy.
YOLO	1. Real time object detection, 2. Finding objects in image grids at parallel.	1. Localization error of bounding boxes, 2. Limited success for detecting small objects, 3. Low recall.
SSD	1. Better balance between rapidity and precision, 2. Very significant in object proposals in their corresponding scales.	1. Limited success for detecting small objects.
YOLOv2	1. Real time object detection, 2. Running significantly faster, 3. High resolution classifier, 4. Multi-scale training.	1. Very complex architecture, 2. Limited success for detecting small objects.
DSSD	1. Significant in detecting small object or context specific objects.	1. Complexity of the model, 2. Slow speed.
DSOD	1. Training object detection networks from scratch with state-of-the-art performance, 2, Great potential on domain different scenarios, Real time object detection, More compact models.	1. Not very good performance.

5 Discussions and conclusions

In this paper, we have focused on identifying the parameters of the image annotation systems based on which we have provided an overview on the *visual content-based* and *users' tags-based image annotation* methods. Then, we have studied the visual content-based images annotation techniques, in particular image segmentation, features extraction and machine/deep learning. Comparisons between the different algorithms that have been illustrated are provided as well.

Overall, the main challenge facing image annotation techniques is the '*semantic gap'* between the low-level visual information captured by the imaging devices and the high-level semantic information perceived by humans [44]. In this regard, many research studies were focused on mining the keyword-keyword and image-keyword relationships. Besides, image annotation process requires a very considerable human intervention when it consists of vast amounts of images. Thereafter, many *Machine-Learning* algorithms were introduced to

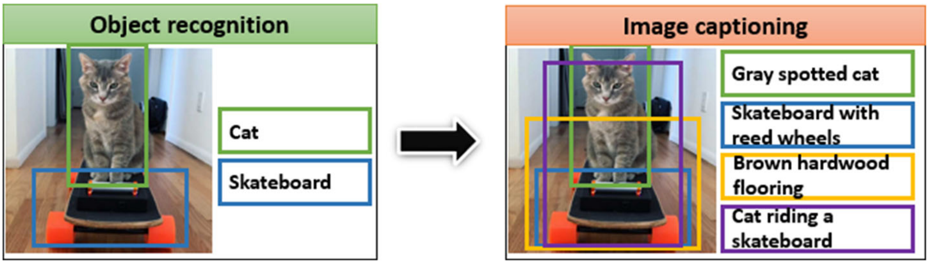

Fig. 22 Image captioning based on the CNN model (example)

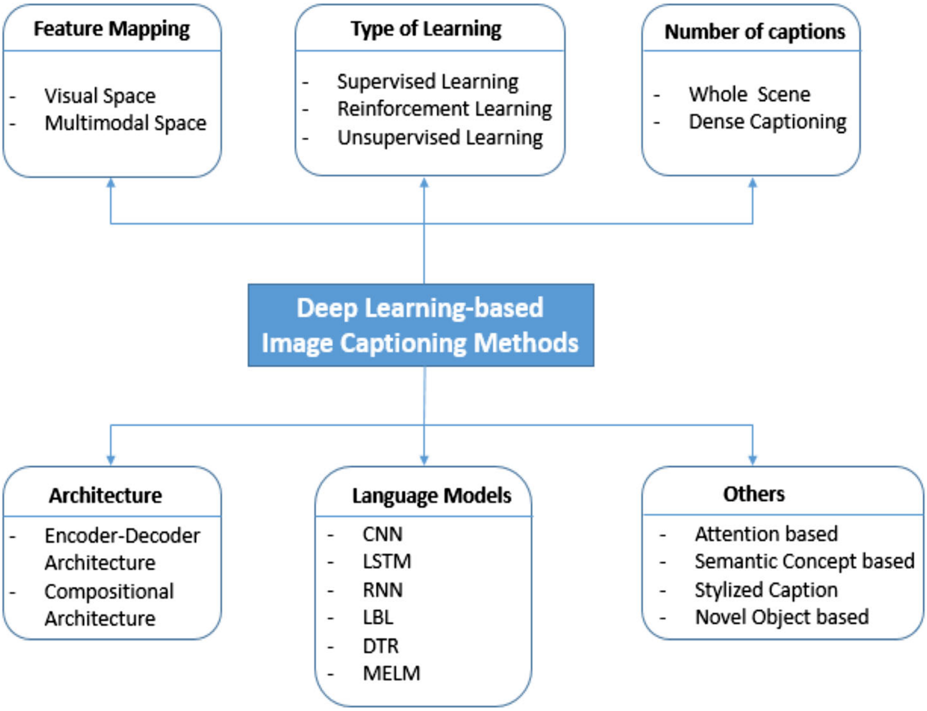

Fig. 23 Image captioning (example)

automate the annotation process and reduce human efforts. However, the main issue in ML is that a bad data representation often degrades the quality of the produced results and leads to lower performance compared with a good data representation. Thus, feature engineering has

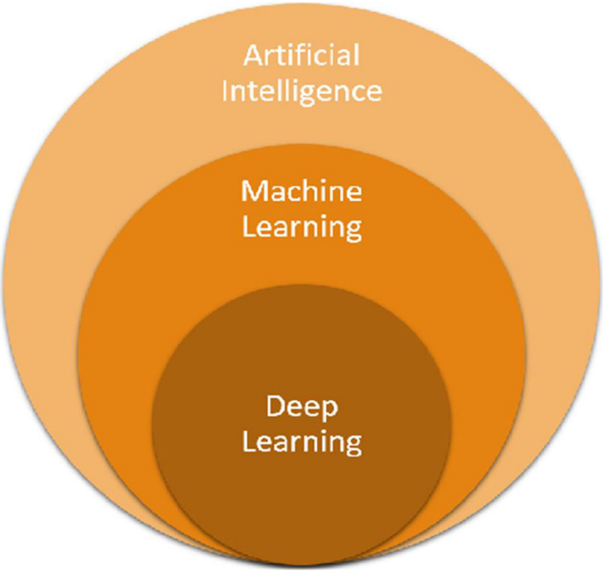

Fig. 24 Subset representation of learning algorithms

been considered an important research direction in ML for a long time. It focuses on extracting deeper features from raw data. This has led to multiple research studies.

In essence, *Deep-Learning* algorithms (DL) provide the opportunity to perform automatic feature extraction. This helps researchers extract discriminative features even with minimal domain information and reduce human efforts. In addition, recent progress in this area show that DL algorithms, in particular CNN, can address the '*semantic gap*' problem. However, there are still numerous open issues for future works related to object detection and image captioning.

Effective Proposal Generation Strategies: To train region detectors, previous approaches needed to manually design anchor boxes, which is a hard way in matching multi-scale objects. As an alternative, recent approaches suggest to use anchor-free methods [155, 279, 339]. However, these methods are costly with the need to be improved. Therefore, designing efficient proposal generation strategies is a very interesting research direction in the future.

Scalable Small Object Detection Strategies: Previous research studies focused on the detection of small objects. In order to improve the localization accuracy of small objects, it becomes interesting to evolve the network architectures. This constitutes in the future a very hot topic in object detection.

Combining the benefits of both one and two-stage detectors: The two-stage detectors, such as R-CNN, Fast R-CNN, Faster R-CNN and Mask R-CNN, follow a densely process in order to achieve as much as reference boxes, which is chronophage and inefficient. To tackle this problem, researchers should avoid redundancy as much as possible while preserving a high accuracy. The one-stage detectors, such as YOLO, YOLOv2, SSD and DSSD, are successfully applied in real-time applications thanks to their fast processing speed that they can achieve. However, the lower accuracy is ever a bottleneck for top precision requirements. Therefore, combining the benefits of both one and two-stage detectors remains an interesting challenge for researchers.

Table 7 Differences between ML and DL

Subject	Machine-learning	Deep-learning
Operating mode	Uses types of automated algorithms that learn to model functions and predict future decisions using the data fed to it.	Interprets data features and their relationships by using neural networks that transmit the relevant information through several data processing steps.
Management	The different algorithms are directed by analysts to examine the different variables in the datasets.	Once they are implemented, the algorithms are generally self-directed for relevant data analysis.
Volume of data	Requires a few thousand of data points used for the analysis.	Requires millions data points used for the analysis.
Output	The output is often a numerical value, such as a score or a classification.	The output may be a score, a free text, an element, or sound, etc.
Feature extraction	Cannot perform automatic feature extraction and needs accurately identified features by human intervention.	Performs automatic feature extraction without the need for human intervention.
Training time	Takes less time to train.	Takes longer to train.
Hardware dependency	Train on CPU.	Requires GPU to train properly.
Hyper-parameter tuning	Limited tuning capabilities.	Can be tuned in various different ways.
Accuracy	Gives lesser accuracy.	Provides high accuracy.

Weakly Supervised Object Detection Methods: Weakly Supervised Object Detection (WSOD) helps use some fully annotated images in order to detect a large amount of no fully annotated ones. Thus, the development of WSOD methods is a significant issue for researchers.

Universal-object Detectors: Explicit domain detectors ever obtain top detection performance on specific domain datasets. Thus, it will be great to have a universal detector that is able to operate on images from different sectors. Indeed, a multi-domain detector can work without prior knowledge on novel domains. Universal-object detector is then a challenging mission.

Unsupervised Object Detection Strategies: Supervised object detection methods are time-consuming and inefficient in training process, hence the need for well-annotated datasets employed as supervision data. However, the annotation of each object's bounding box in big databases is costly, hard and impractical. Therefore, the development of automatic annotation strategies to lighten the human annotation intervention is a promising solution for an unsupervised object detection. Unsupervised object detection is a very interesting research direction in the future.

Multi-source Information Assistance Strategies: Thanks to the popularity of social media and the progress of big data processing technologies, multi-source information have become easy to be accessed. Numerous social networks help host both images and descriptions associated with them in textual form. This type of information can facilitate the detection task. Therefore, multi-source information assistance is an emerging research direction for objet detection in the future.

3D datasets and Object Detection Strategies: With the emergence of 3D sensors applications, deeper additional information can be used for a better understanding of the content of the 2D and real-world images. Therefore, there is a great need for large-scale 3D image datasets as well as techniques that aim at correctly detecting 3D bounding boxes around objects.

Effective Image Captioning Strategies: Object detection has achieved in important success in recent years. However, the detection of the attributes of the objects as well as the relationships between them is a still open topic that requires lot efforts to achieve high-quality image captions. Besides, designing sophisticated language generation models is an interesting research direction in the future seeing that the accuracy of the generated captions mainly depends on the quality of their syntax. Finally, supervised learning requires a vast amount of tagged data for training. Thus, it will be interesting to rely on unsupervised and reinforcement learning in the future.

References

1. Abdel-Hamid O, Mohamed AR, Jiang H, Deng L, Penn G, Yu D (2014) Convolutional neural networks for speech recognition. IEEE/ACM transactions on audio, speech, and language processing. IEEE/ACM 22(10): 1533–1545
2. Abioui H, Idarrou A, Bouzit A, Mammass D: Review: Automatic Image Annotation for Semantic Image Retrieval. In: Proceedings of the 6th International Conference on Image and Signal Processing (ICISP), pp. 129-137. Springer, Cherbourg, France (2018)
3. Abo-Zahhad M, Gharieb RR, Ahmed SM, Donkol AAEB (2014) Edge detection with a preprocessing approach. Journal of Signal and Information Processing (*JSIP*) 5(4):123–134
4. Adebayo S, McLeod K, Tudose I, Osumi-Sutherland D, Burdett T, Baldock R, Parkinson H (2016) PhenoImageShare: an image annotation and query infrastructure. Journal of Biomedical Semantics 7(1): 35–44

5. Ajala Funmilola A, Oke OA, Adedeji TO, Alade OM, Adewusi E (2012) A: fuzzy k-means clustering algorithm for medical image segmentation. Journal of Information Engineering and Applications 2(6):21–32
6. Akbulut Y, Sengur A, Guo Y, Smarandache F (2017) NS-k-NN: Neutrosophic set-based k-nearest Neighbors classifier. Symmetry 9(9):179
7. Alham N. K, Li M, Liu Y, Hammoud S, Ponraj M: A distributed SVM for scalable image annotation. In: Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 2655-2658. IEEE, Shanghai, China (2011)
8. Anees V M, Kumar G S, Sreeraj M: Automatic image annotation using SURF descriptors. In: Proceedings of the 2012 Annual IEEE India Conference (INDICON), pp. 920-924. IEEE, Kochi, India (2012)
9. Aneja J, Deshpande A, Schwing A G: Convolutional image captioning. In: Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5561–5570. IEEE, Honolulu, HI, USA (2017)
10. Angelina S, Suresh L P, Veni S K: Image segmentation based on genetic algorithm for region growth and region merging. In: Proceedings of the 2012 IEEE International Conference on Computing, Electronics and Electrical Technologies (ICCEET), pp. 970-974. IEEE, Kumaracoil, India (2012)
11. Anjna EA, Er RK (2017) Review of image segmentation technique. Int J Adv Res Comput Sci 8(4):36–39
12. Appels R, Nystrom-Persson J, Keeble-Gagnere G (2014) Advances in genome studies in plants and animals. Functional et Integrative Genomics Springer 14(1):1–9
13. Arellano G, Sucar L E, Morales E F: Automatic image annotation using multiple grid segmentation. In: Proceedings of the Mexican International Conference on Artificial Intelligence (MICAI), pp. 278-289. Springer, Pachuca (2010)
14. Arun K. Pujari, Data mining techniques-a reffrence book ,pg. no.-114-147 (2013)
15. Atlam HF, Attiya G, El-Fishawy N (2017) Integration of color and texture features in CBIR system. Int J Comput Appl 164(3):23–29
16. Ayadi Y, Amous I, Gargouri F (2013) Toward an automatic annotation approach based on ontological enrichment for advanced research. International Journal of Engineering et Technology (IJET-IJENS) 13(2): 80–89
17. Badrinarayanan V, Kendall A, Cipolla R: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. CoRR, abs/1511.00561 (2015)
18. Bay H, Tuytelaars T, Van Gool L: Surf: Speeded up robust features. In: Proceedings of the 9th European Conference on Computer Vision (ECCV), pp. 404– 417. Springer, Graz, Austria (2006)
19. Belkhatir M (2009) An operational model based on knowledge representation for querying the image content with concepts and relations. Multimedia Tools and Applications Springer 43(1):1–23
20. Bell S., Upchurch P, Snavely N, Bala K: Material recognition in the wild with the materials in context database. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3479-3487. IEEE, Boston, MA, USA (2015)
21. Bergeaud F, Mallat S: Matching pursuit of images. In: Proceedings of the 1995 IEEE International Conference on Image Processing (ICIP), pp. 53-56. IEEE, Washington, DC, USA (1995)
22. Bhatt H S, Bharadwaj S, Singh R, Vatsa M: On matching sketches with digital face images. In: Proceedings of the 4th International Conference on Biometrics Theory Applications and Systems (BTAS), pp. 1-7. IEEE, Washington, DC, USA (2010)
23. Bhende P, Cheran, AN.: Content based image retrieval in Medical Imaging. International Journal of Computational Engineering and Research. (*IJCER*). **3**(8), 10-15 (2013)
24. Blei D M, Jordan M I: Modeling annotated data. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 127-134. ACM, Toronto, Canada (2003)
25. Bobade KB, Jagtap SV (2014) Automatic image annotation by classification using SIFT features. International Journal of Scientific Research Engineering & Technology 3(3):713–720
26. Bouchakwa M, Ayadi Y, Amous I: Modeling the semantic content of the socio-tagged images based on the extended conceptual graphs formalism. In: Proceedings of the 14th International Conference on Advances in Mobile Computing and MultiMedia (MOMM), pp. 35-39. ACM, Singapore (2016)
27. Bouchakwa M, Ayadi Y, Amous I: Semantic Pattern-based Automatic Annotation Process of Images Shared on Social Networks. In: Proceedings of the 30th IBIMA Conference (IBIMA), pp. 19. Madrid, Spain (2017)
28. Bouchakwa M, Ayadi Y, Amous I: Multi-level diversification approach of semantic-based image retrieval results. Progress in Artificial Intelligence (PAI). 1-30 (2019)
29. Boutell MR, Luo J, Shen X, Brown CM (2004) Learning multi-label scene classification. Pattern recognition Elsevier science 37(9):1757–1771
30. Bovik AC, Clark M, Geisler WS (1990) Multichannel texture analysis using localized spatial filters. IEEE transactions on pattern analysis machine intelligence. (TPAMI). IEEE 12(1):55–73

31. Boykov Y Y, Jolly M P: Interactive graph cuts for optimal boundary et region segmentation of objects in ND images. In: Proceedings of the 8th IEEE International Conference on Computer Vision (ICCV), pp. 105-112. IEEE, Vancouver, Canada (2001)
32. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman&Hall (Wadsworth). Monterey, California, USA
33. Cannon RL, Dave JV, Bezdek JC, Trivedi MM (1986) Segmentation of a thematic mapper image using the fuzzy c-means clusterng algorthm. IEEE transactions on geoscience and remote sensing (TGRS). IEEE 24(3):400–408
34. Carson C, Belongie S, Greenspan H, Malik J (2002) Blobworld: image segmentation using expectation-maximization and its application to image. IEEE Transactions on Pattern Analysis and Machine Intelligence IEEE 24(8):1026–1038
35. Chakraborty A, Duncan JS (1999) Game-theoretic integration for image segmentation. IEEE transactions on pattern analysis and machine intelligence (PAMI). IEEE 21(1):12–30
36. Chan TF, Vese LA (2001) Active contours without edges. IEEE transactions on image processing (TIP). IEEE 10(2):266–277
37. Chang T, Kuo CC (1993) Texture analysis and classification with tree-structured wavelet transform. IEEE transactions on image processing (TIP). IEEE 2(4):429–441
38. Chapelle O, Haffner P, Vapnik VN (1999) Support vector machines for histogram-based image classification. IEEE Transactions on Neural Networks IEEE 10(5):1055–1064
39. Chathurani N W U D, Geva S, Chandran V, Cynthujah V: An effective content based image retrieval system based on global representation and multi-level searching. In: Proceedings of the 10th International Conference on Industrial and Information Systems (ICIIS), pp. 158-163. IEEE, Peradeniya, Sri Lanka (2015)
40. Chaudhuri BB, Sarkar N (1995) Texture segmentation using fractal dimension. IEEE transactions on pattern analysis and machine intelligence (TPAMI). 17:1, 72–IEEE, 77
41. Chen Y, Wang JZ (2002) A region-based fuzzy feature matching approach to content based image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence IEEE 24(9):1252–1267
42. Chen Y, Wang JZ (2004) Image categorization by learning and reasoning with regions. The Journal of Machine Learning Research (JMLR) ACM 5:913–939
43. Xinlei Chen and C Lawrence Zitnick.: Mind's eye: A recurrent visual representation for image caption generation. In: Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2422–2431. IEEE, Boston, MA, USA (2015)
44. Chen X, Yuan X, Yan S, Tang J, Rui Y, Chua T S: Towards multi-semantic image annotation with graph regularized exclusive group lasso. In: Proceedings of the 19th ACM International Conference on Multimedia (MM), pp. 263-272. ACM, Scottsdale, AZ, USA (2011)
45. Chen L C, Papandreou G, Kokkinos I, Murphy K, Yuille A L: Semantic image segmentation with deep convolutional nets and fully connected crfs. CoRR, abs/1412.7062 (2014)
46. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In: IEEE transactions on pattern analysis and machine intelligence (TPAMI). IEEE 40(4):834–848
47. Cheng Q, Zhang Q, Fu P, Tu C, Li S (2018) A survey and analysis on automatic image annotation. Pattern Recogn 79(2018):242–259
48. Chengjian S, Zhu S, Shi Z: Image annotation via deep neural network. In: Proceedings of the 14th IAPR International Conference on Machine Vision Applications (MVA), pp. 518-521. IEEE, Tokyo, Japan (2015)
49. Choi D, Kim P: Automatic image annotation using semantic text analysis. In: Proceedings of the 7th International Conference on Availability, Reliability, and Security (ARES), pp. 479-487. Springer, Prague, Czech Republic (2012)
50. Clerc M, Kennedy J (2002) The particle swarm-explosion, stability, and convergence in a multidimensional complex space. IEEE transactions on evolutionary computation (TEVC). IEEE 6(1):58–73
51. Cooper L, Walls RL, Elser J, Gandolfo MA, Stevenson DW, Smith B, Hiss M (2012) The plant ontology as a tool for comparative plant anatomy and genomic analyses. Plant Cell Physiol 54(2):1–23
52. Cross GR, Jain AK (1983) Markov random field texture models. IEEE transactions on pattern analysis and machine intelligence (TPAMI). IEEE 5(1):25–39
53. Cusano C, Ciocca G, Schettini R: Image annotation using SVM. In: International Society for Optics and Photonics (SPIE), pp. 330-339 (2003)
54. Dai J, Li Y, He K, Sun J: R-fcn: Object detection via region-based fully convolutional networks. In: Proceedings of the 30th Advances in Neural Information Processing Systems (NIPS), pp. 379-387. Barcelona, Spain (2016)

55. Dai B, Fidler S, Urtasun R, Lin D: Towards Diverse and Natural Image Descriptions via a Conditional GAN. In: Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2989–2998. IEEE, Honolulu, HI, USA (2017)
56. Dalal N, Triggs B: Histograms of Oriented Gradients for Human Detection. In: Proceedings of the 15th Computer Vision and Pattern Recognition (CVPR), pp. 886-893. IEEE, San Diego, CA, USA (2005)
57. Daugman JG (1985) Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. Journal of the Optical Society of America A (JOSA A) 2(7):1160–1169
58. Deng Y, Manjunath BS (2001) Unsupervised segmentation of color-texture regions in images and video. IEEE transactions on pattern analysis and machine intelligence (TPAMI). IEEE 23(8):800–810
59. Deng J, Dong W, Socher R, Li L J, Li K, Fei-Fei L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248-255. IEEE, Miami, FL, USA (2009)
60. Derin H, Elliott H, Cristi R, Geman D (1984) Bayes smoothing algorithms for segmentation of binary images modeled by Markov random fields. IEEE transactions on pattern analysis and machine intelligence (PAMI). IEEE 6(6):–707, 720
61. Dharani T, Aroquiaraj I L: A survey on content based image retrieval. In: Proceedings of the 2013 IEEE International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), pp. 485-490. IEEE, Tamilnadu, India (2013)
62. Dimitrovski I, Kocev D, Loskovska S, Dzeroski S: Detection of Visual Concepts and Annotation of Images Using Predictive Clustering Trees. In : CLEF (Notebook Papers/LABs/Workshops), pp. 1-10 (2010)
63. Domingos P, Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. Machine learning Springer 29(2-3):103–130
64. Erhan D, Szegedy C, Toshev A, Anguelov D: Scalable object detection using deep neural networks. In: Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2147-2154. IEEE, Columbus, OH, USA (2014)
65. Fan J, Gao Y, Luo H, et Xu G: Automatic image annotation by using concept-sensitive salient objects for image content representation. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 361-368. ACM, Sheffield, United Kingdom (2004)
66. Fang H, Gupta S, Iandola F, Srivastava R K, Deng L, Dollár P, Lawrence Zitnick C: From captions to visual concepts and back. In: Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1473-1482. IEEE, Boston, MA, USA (2015)
67. Farhadi A, Hejrati M, Sadeghi M A, Young P, Rashtchian C, Hockenmaier J, Forsyth D: Every picture tells a story: Generating sentences from images. In: Proceedings of the 11th European Conference on Computer Vision (ECCV), pp. 15-29. Springer, Heraklion, Crete, Greece (2010)
68. Feng H, Chua T S: A bootstrapping approach to annotating large image collection. In: Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 55-62. ACM, Berkeley, California (2003)
69. Feng S L, Manmatha R, Lavrenko V: Multiple Bernoulli relevance models for image and video annotation. In: Proceedings of the 2004 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1002-1009. IEEE, Washington, DC, USA, (2004)
70. Figueiredo J C, Neto F G M, de Paula I C: Contour-based feature extraction for image classification and retrieval. In: Proceedings of the 35th International Conference of the Chilean Computer Science Society (SCCC), pp. 1-7. IEEE, Valparaiso, Chile (2016)
71. Franco-Lopez H, Ek AR, Bauer ME (2001) Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. Remote sensing of Environment Elsevier science 77(3): 251–274
72. Fu C Y, Liu W, Ranga A, Tyagi A, Berg A C: Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659* (2017)
73. Gan C, Gan Z, He X, Gao J, Deng L: Stylenet: Generating attractive visual captions with styles In: Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3137–3146. IEEE, Honolulu, HI, USA (2017)
74. Gao YY, Yi-Xin YIN, Uozumi T (2010) A hierarchical image annotation method based on SVM and semi-supervised EM. Acta Automatica Sinica Elsevier science 36(7):960–967
75. Garcia-Garcia A, Orts-Escolano S., Oprea S, Villena-Martinez V, Garcia-Rodriguez J: A review on deep learning techniques applied to semantic segmentation. CoRR, abs/ 1704.06857 (2017)
76. Geman S, Geman D: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). IEEE **20**(6-5), 721-741 (1984)

77. Ghahabi O, Hernando Pericás FJ (2018) Restricted Boltzmann machines for vector representation of speech in speaker recognition. Computer Speech and Language Elsevier science 47:16–29
78. Ghoshal A, Ircing P, Khudanpur S: Hidden Markov models for automatic annotation and content-based retrieval of images and video. In: Proceedings of the 28th annual International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 544-551. ACM Salvador, Brazil (2005)
79. Girshick R: Fast r-cnn. In: Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV), pp. 1440-1448. IEEE, Santiago, Chile (2015)
80. Girshick R, Donahue J, Darrell T, Malik J: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580-587. IEEE, Columbus, OH, USA (2014)
81. Goh K S, Chang E Y, Li B: Using one-class and two-class SVMs for multiclass image annotation. IEEE Transactions on Knowledge and Data Engineering (TKDE). IEEE **17**(10), 1333-1346 (2005)
82. Göksu Ö, Aptoula E: Content based image retrieval of remote sensing images based on deep features. In: Proceedings of the 26th Signal Processing and Communications Applications Conference (SIU), pp. 1-4. IEEE, Izmir, Turkey (2018)
83. Gong T, Li S, Tan C L: A semantic similarity language model to improve automatic image annotation. In: Proceedings of the 22nd International Conference on Tools with Artificial Intelligence (ICTAI), pp. 197-203. IEEE, Arras, France (2010)
84. Gong Y, Jia Y, Leung T, Toshev A, Ioffe S: Deep convolutional ranking for multilabel image annotation. CoRR, abs/1402.1128 (2013)
85. Gong Y, Wang L, Hodosh M, Hockenmaier J, Lazebnik S: Improving image-sentence embeddings using large weakly annotated photo collections. In: Proceedings of the 13th European Conference on Computer Vision (ECCV), pp. 529-545. Springer, Zurich, Switzerland (2014)
86. Grady L: Random walks for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). IEEE **28**(11), 1768-1783 (2006)
87. Grady L, Schwartz E L: Isoperimetric graph partitioning for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). IEEE **28**(3), 469-475 (2006)
88. Gu J, Wang G, Cai J, Chen T: An empirical study of language cnn for image captioning. In: Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), pp. 1231–1240. IEEE, Venice, Italy (2017)
89. Guha S, Rastogi R, Shim K (1998) CURE: an efficient clustering algorithm for large databases. ACM Sigmod Record ACM 27(2):73–84
90. Guillaumin M, Mensink T, Verbeek J, Schmid C: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: Proceedings of the 12th International Conference on Computer Vision (*ICCV*), pp. 309-316. IEEE, Kyoto, Japan (2009)
91. Guru D S, Sharath Y H, Manjunath S: Texture features and KNN in classification of flower images. International Journal of Computer Applications (IJCA), Special Issue on Recent Trends in Image Processing and Pattern Recognition. (1), 21-29 (2010)
92. Halaschek-Wiener C, Golbeck J, Schain A, Grove M, Parsia B, Hendler J: Photostuff: An image annotation tool for the semantic web. In: Proceedings of the 4th International Semantic Web Conference (ISWC), pp. 6-10. Springer, Galway, Ireland (2005)
93. Hambali H A, Abdullah S L S, Jamil N, Harun H: Fruit Classification using Neural Network Model. Journal of Telecommunication, Electronic and Computer Engineering (JTEC). 9(1-2), 43-46 (2017)
94. Han Y, Qi X: A complementary svms-based image annotation system. In: Proceedings of the 2005 IEEE International Conference on Image Processing (ICIP), pp. 1185-1188. IEEE, Genoa, Italy (2005)
95. Hanbury A: A survey of methods for image annotation. Journal of Visual Languages & Computing (JVLC). Elsevier science **19**(5), 617-627 (2008)
96. Haralick RM (1979) Statistical and structural approaches to texture. Proceedings of the IEEE IEEE 67(5): 786–804
97. Harzallah H, Jurie F, Schmid C: Combining efficient object localization and image classification In : Proceedings of the 12th IEEE International Conference on Computer Vision (ICCV), pp. 237-244. IEEE, Kyoto, Japan (2009)
98. Hastings S, Oster S, Langella S, Kurc TM, Pan T, Catalyurek UV, Saltz JH (2005) A grid-based image archival and analysis system. Journal of the American medical informatics association (JAMIA). Elsevier science 12(3):286–295
99. He X J, Zhang Y, Lok T M, Lyu M R: A new feature of uniformity of image texture directions coinciding with the human eyes perception. In: Proceedings of the 2nd International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 727-730. Springer, Changsha, China (2005)
100. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916

101. He K, Gkioxari G, Dollár P, Girshick R: Mask r-cnn. In: Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), pp. 2980-2988. IEEE Venice, Italy (2017)
102. Hermanto A, Adji T B, Setiawan N A: Recurrent neural network language model for English-Indonesian Machine Translation: Experimental study. In: Proceedings of the 2015 International Conference on Science in Information Technology (ICSITech), pp. 132-136. IEEE, Yogyakarta, Indonesia (2015)
103. Hiremath P S, Pujari J: Content based image retrieval using color, texture and shape features. In: Proceedings of the 15th International Conference on Advance Computing and Communications (ADCOM), pp. 780-784. IEEE, Guwahati, Assam (2007)
104. Hodosh M, Young P, Hockenmaier J (2013) Framing image description as a ranking task: data, models and evaluation metrics. J Artif Intell Res 47(1):853–899
105. Hollink L, Schreiber A T, Wielemaker J, Wielinga B J: Semantic annotation of image collections. p. 8 (2003)
106. Hollink L, Nguyen G, Schreiber G, Wielemaker J, Wielinga B, Worring M: Adding spatial semantics to image annotations. In: Proceedings of the 4th International Workshop on Knowledge Markup and Semantic Annotation at ISWC, pp.31-40. Hiroshima, Japan (2004)
107. Horvat M, Grbin A, Gledec G (2013) Labeling and retrieval of emotionally-annotated images using WordNet. International Journal of Knowledge-based and Intelligent Engineering Systems ACM 17(2): 157–166
108. Hossain MD, Sohel F, Shiratuddin MF, Laga H (2019) A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CSUR) 51(6):118–154
109. Huang Y F, Lu H Y: Automatic image annotation using multi-object identification. In: Proceedings of the 4th Pacific-Rim Symposium on Image and Video Technology (PSIVT), pp. 386-392. IEEE, Singapore (2010)
110. Huang J, Kumar S R, Mitra M, Zhu W J, Zabih R: Image indexing using color correlograms. In: Proceedings of the 1997 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 762-768. IEEE, San Juan, Puerto Rico, USA (1997)
111. Huang J, Liu H, Shen J, Yan S: Towards efficient sparse coding for scalable image annotation. In: Proceedings of the 21st ACM International Conference on Multimedia (MM), pp. 947-956. ACM, Barcelona, Spain (2013)
112. Im D H, Park G D: STAG: semantic image annotation using relationships between tags. In: Proceedings of the 2013 International Conference on Information Science and Applications (ICISA), pp. 1-2. IEEE, Suwon, South Korea (2013)
113. Im DH, Park GD (2015) Linked tag: image annotation using semantic relationships between image tags. Multimedia Tools and Applications Springer 74(7):2273–2287
114. Islam M M, Zhang D, Lu G: A geometric method to compute directionality features for texture images. In: Proceedings of the 2008 IEEE International Conference on Multimedia and Expo (ICME), pp. 1521–1524. IEEE, Hannover, Germany (2008)
115. Islam M M, Zhang D, Lu G: Automatic categorization of image regions using dominant color based vector quantization. In: Proceedings of the 2008 IEEE Digital Image Computing: Techniques and Applications (DICTA), pp. 191–198. IEEE, Canberra, Australia (2008)
116. Jaderberg M, Simonyan K, Zisserman A: Spatial transformer networks. In: Proceedings of the Advances in Neural Information Processing Systems (NIPS), pp. 2017-2025. Montréal CANADA (2015)
117. Jain AK, Vailaya A (1996) Image retrieval using color and shape. Pattern recognition Elsevier science 29(8):1233–1244
118. Jau-Ling S, Ling-Hwei C: Color image retrieval based on primitives of color moments. In: Proceedings of the 5th International Conference on Advances in Visual Information Systems (VISUAL), pp. 88-94. Springer, Hsin Chu, Taiwan (2002)
119. Jeon J, Lavrenko V, Manmatha R: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 119-126. ACM, Toronto, Canada (2003)
120. Jeong J W, Hong H K, Lee D H: i-TagRanker: an efficient tag ranking system for image sharing and retrieval using the semantic relationships between tags. Multimedia Tools and Applications. Springer **62**(2), 51-478 (2013)
121. Ji Q, Zhang L, Li Z: KNN-based Image Annotation by Collectively Mining Visual and Semantic Similarities. Transactions on Internet & Information Systems (KSII). **11**(9), 4476-4490 (2017)
122. Jia X, Gavves E, Fernando B, Tuytelaars T: Guiding the long-short term memory model for image caption generation. In: Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV), pp. 2407–2415. IEEE, Santiago, Chile (2015)

123. Jiang Z, He J, Guo P: Feature data optimization with LVQ technique in semantic image annotation. In: Proceedings of the 10th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 906-911. IEEE, Cairo, Egypt (2010)
124. Jiawei H, Michheline K: Data mining concepts and techniques-a reffrence book ,pg. no.-383-422
125. Jin Y, Khan L, Wang L, Awad M: Image annotations by combining multiple evidence et wordnet. In: Proceedings of the 13th Annual ACM International Conference on Multimedia (MM), pp. 706-715. ACM, Singapore (2005)
126. Jin J, Fu K, Cui R, Sha F, Zhang C: Aligning where to see and what to tell: image caption with region-based attention and scene factorization. *arXiv preprint arXiv:1506.06272* (2015)
127. Jing F, Li M, Zhang L, Zhang H J, Zhang B: Learning in region-based image retrieval. In: Proceedings of the 2nd International Conference on Image and Video Retrieval (CIVR), pp. 206-215. Springer, Urbana-Champaign, IL, USA (2003)
128. Joachims T: Optimizing search engines using clickthrough data. In: Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 133-142. ACM, Edmonton, Alberta, Canada (2002)
129. John G H, Langley P: Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI), pp. 338-345. ACM, Montréal, Canada (1995)
130. Johnson J, Karpathy A, Fei-Fei L: Densecap: Fully convolutional localization networks for dense captioning. In: Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4565-4574. IEEE, Las Vegas, NV, USA (2016)
131. Kalafi EY, Tan WB, Town C, Dhillon SK (2016) Automated identification of monogeneans using digital image processing and K-nearest neighbor approaches. BMC bioinformatics 17(19):511
132. Kamdi S, Krishna R K: Image segmentation and region growing algorithm. International Journal of Computer Technology and Electronics Engineering (IJCTEE). **2**(1), 103-107 (2012)
133. Karoui I, Fablet R, Boucher JM, Augustin JM (2010) Variational region-based segmentation using multiple texture statistics. IEEE Transactions on Image Processing (TIP) 19(12):3146–3156
134. Karpathy A, Fei-Fei L: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3128–3137. IEEE, Boston, MA, USA (2015)
135. Karpathy A, Joulin A, Fei-Fei L.: Deep fragment embeddings for bidirectional image sentence mapping. In: Proceedings of the 29th Advances in Neural Information Processing Systems (NIPS), pp. 1889–1897. Montreal, Quebec, Canada (2014)
136. Karypis G, Han EH, Kumar V (1999) Chameleon: hierarchical clustering using dynamic modeling. Computer IEEE 32(8):68–75
137. Kass M, Witkin A, Terzopoulos D (1988) Snakes: active contour models. International Journal of Computer Vision Springer 1(4):321–331
138. Kaya Y, Kayci L (2014) Application of artificial neural network for automatic detection of butterfly species using color and texture features. The Visual Computer Elsevier science 30(1):71–79
139. Kendall A, Badrinarayanan V, Cipolla R: Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. CoRR, abs/1511.02680 (2015)
140. Kennedy J, Eberhart R.: Particle swarm optimization. In: Proceedings of the 5th IEEE International Conference on Neural Networks (ICANN), pp. 1942-1948. IEEE, Paris, France (1995)
141. Khan A, Deep S, Li J P, Kumar K, Shaikh R A, Hasan F: Vision prehension with CBIR for cloud robo. In: Proceedings of the 11th International Computer Conference on Wavelet Actiev Media Technology and Information Processing (ICCWAMTIP), pp. 293-296. IEEE, China, Sichuan Province (2014)
142. Kiros, R., Szepesvári, C.: Deep representations and codes for image auto-annotation. In: Proceedings of 26th Annual Conference on Neural Information Processing Systems (NIPS), pp. 908-916. Lake Tahoe, Nevada, USA (2012)
143. Kiros R, Salakhutdinov R, Zemel R: Multimodal neural language models. In: Proceedings of the 31st International Conference on Machine Learning (ICML), pp. 595–603. Beijing, China (2014)
144. Kiros J R, Salakhutdinov R, Zemel R: Unifying visual-semantic embeddings with multimodal neural language models. In: Proceedings of the 28th Workshop on Neural Information Processing Systems (NIPS). Montreal, Quebec, Canada (2014)
145. Krishnan KB, Ranga SP (2017) Guptha. N: A Survey on Different Edge Detection Techniques for Image Segmentation Indian Journal of Science and Technology 10(4):1–8
146. Krizhevsky A, Sutskever I, Hinton G E: Imagenet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems (NIPS), pp. 1097-1105 (2012)

147. Ksibi A, Ammar A B, Amar C B: Effective concept detection using second order co-occurence flickr context similarity measure socfcs. In: Proceedings of the 10th International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 1-6. IEEE, Annecy, France (2012)

148. Kulkarni G, Premraj V, Dhar S, Li S, Choi Y, Berg A C, Berg T L.: Baby talk: Understanding and generating image descriptions. In: Proceedings of the 24th Computer Vision and Pattern Recognition (CVPR), pp. 1601-1608. IEEE, Colorado Springs, CO, USA (2011)

149. Kumar K K: CBIR: Content based image retrieval. In: Proceedings of the 2010 National Conference on Recent Trends in information/ Network Security (NCRTNS), pp. 36-43 (2010)

150. Kuroda K, Hagiwara M (2002) An image retrieval system by impression words and specific object names–IRIS. Neurocomputing Elsevier science 43(1-4):259–276

151. Kurtz C, Rubin D L: Utilisation de relations ontologiques pour la comparaison d'images décrites par des annotations sémantiques, In: Proceedings of the 14th Conference on Knowledge Extraction and Management (EGC), pp. 609-614. Rennes (2014)

152. Kwitt, R., Vasconcelos, N., Rasiwasia, N., Uhl, A., Davis, B., Häfner, M., Wrba, F.: Endoscopic image analysis in semantic space. Medical Image Analysis (MIA). **16**(7), 1415-1422 (2012)

153. Laine A, Fan J: Texture classification by wavelet packet signatures. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). IEEE **15**(11), 1186-1191 (1993)

154. Lavrenko V, Manmatha R, Jeon J: A model for learning the semantics of pictures. In: Proceedings of the 16th International Conference on Neural Information Processing Systems (NIPS), pp. 553-560. ACM, Whistler, British Columbia, Canada (2003)

155. Law H, Deng J: Cornernet: Detecting objects as paired keypoints. In: Proceedings of the 15th European Conference on Computer Vision (ECCV), pp. 734-750. Springer, Munich, Germany (2018)

156. Lazebnik S, Schmid C, Ponce J: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2169-2178. IEEE, New York, NY, USA (2006)

157. Leacock C, Chodorow M (1998) Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database ACM 49(2):265–283

158. Lei Y, Wong W, Liu W, Bennamoun M: An HMM-SVM-based automatic image annotation approach. In: Proceedings of the 10th Asian Conference on Computer Vision (ACCV), pp. 115-126. Springer, Queenstown, New Zealand (2010)

159. Levine M: Vision in Man and Machine, McGraw-Hill (1985)

160. Lew M S, Sebe N, Djeraba C, Jain R: Content-based multimedia information retrieval: state of the art and challenges. ACM Transactions on Multimedia Computing, Communications and Applications (TOMM). ACM **2**(1), 1–19 (2006)

161. Li B, Goh K: Confidence-based dynamic ensemble for image annotation and semantics discovery. In: Proceedings of the 11th ACM International Conference on Multimedia (MM), pp. 195-206. ACM, Berkeley, CA, USA (2003)

162. Li J, Wang J Z, Wiederhold G: IRM: Integrated region matching for image retrieval. In: Proceedings of the 8th ACM international conference on Multimedia, pp. 147-156. ACM, Marina del Rey, California, USA (2000)

163. Li S, Kulkarni G, Berg T L, Berg A C, Choi Y: Composing simple image descriptions using web-scale n-grams. In: Proceedings of the 15th Conference on Computational Natural Language Learning (CoNLL), pp. 220-228. ACM, Portland, Oregon (2011)

164. Li T, Cheng B, Ni B, Liu G, Yan S: Multitask low-rank affinity graph for image segmentation and image annotation. ACM Transactions on Intelligent Systems and Technology (TIST). **7**(4), 1-18 (2016)

165. Li Y D, Hao Z B, Lei H: Survey of convolutional neural network. International Journal of Computer Applications (IJCA). **36**(9), 2508-2515 (2016)

166. Lin D: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning (LCML), pp. 296-304. ACM, San Francisco, CA, USA (1998)

167. Lingutla NT, Preece J, Todorovic S, Cooper L, Moore L, Jaiswal P (2014) AISO: annotation of image segments with ontologies. Journal of Biomedical Semantics Springer 5(1):50–54

168. Liu Y, Zhang D, Lu G, Ma W Y: Region-based image retrieval with perceptual colors. In: Proceedings of the 5th Pacific-Rim Conference on Multimedia (PCM), pp. 931-938. Springer, Tokyo, Japan (2004)

169. Liu Y, Zhang D, Lu G, Ma WY (2007) A survey of content-based image retrieval with high-level semantics. Pattern Recognition Elsevier science 40(1):262–282

170. Liu D, Hua X S, Wang M, Zhang H J: Image retagging. In: Proceedings of the 18th ACM International Conference on Multimedia (MM), pp. 491-500. ACM, Firenze, Italy (2010)

171. Liu W, Ji R, Li S: Towards 3d object detection with bimodal deep boltzmann machines over rgbd imagery. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3013-3021. IEEE, Boston, MA, USA (2015)

172. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, Berg A C: Ssd: Single shot multibox detector. In: Proceedings of the 14th European Conference on Computer Vision (ECCV), pp. 21-37. Springer, Cham (2016)
173. Long F, Zhang H, Feng D D: Fundamentals of content-based image retrieval. In: Proceedings of 2003 International Conference on Multimedia Information Retrieval and Management (MIRM), pp. 1-26. Springer, Berlin, Heidelberg (2003)
174. Long J, Shelhamer E, Darrell T: Fully convolutional networks for semantic segmentation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431-3440. IEEE, Boston, MA, USA (2015)
175. Long J, Shelhamer E, Darrell T: Fully convolutional networks for semantic segmentation. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431-3440. IEEE, Las Vegas, NV, USA (2015)
176. Low W C, Chua T S: Colour-based relevance feedback for image retrieval. In: Proceedings of the 1998 IEEE International Workshop on Multi-Media Database Management Systems, pp. 116-123. IEEE, Dayton, OH, USA (1998)
177. Lowe D G: Object recognition from local scale-invariant features. In: Proceedings of the 7th IEEE International Conference on Computer Vision (*ICCV*), pp. 1150–1157. IEEE, Kerkyra, Corfu, Greece (1999)
178. Lu CS, Chung PC, Chen CF (1997) Unsupervised texture segmentation via wavelet transform. Pattern Recognition Elsevier science 30(5):729–742
179. Lu H, Zheng Y, Xue X, Zhang Y: Content and context-based multi-label image annotation. In: Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR), pp. 61-68. IEEE, Miami, FL, USA (2009)
180. Lu J, Xiong C, Parikh D, Socher R: Knowing when to look: Adaptive attention via A visual sentinel for image captioning. In: Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3242–3250. IEEE, Honolulu, HI, USA (2017)
181. Magesh N, Thangaraj P: Semantic image retrieval based on ontology and SPARQL query. In: Proceedings of the 2nd International Conference on Advanced Computer Technology (ICACT), pp. 12-16. IEEE, Gangwon-Do, Korea (2011)
182. Makadia A, Pavlovic V, Kumar S: A new baseline for image annotation. In: Proceedings of the 10th European Conference on Computer Vision (ECCV), pp. 316-329. Springer, Marseille, France (2008)
183. Mallat S G: Multifrequency channel decompositions of images and wavelet models. IEEE Transactions on Acoustics, Speech, and Signal Processing. IEEE **37**(12), 2091-2110 (1989)
184. Mallat S, Zhang Z: Matching pursuit with time-frequency dictionaries. IEEE Transactions on Signal Processing (TSP). IEEE **41**(12), 3397-3415 (1993)
185. Manjunath B S, Ohm J R, Vasudevan V V, Yamada A: Color and texture descriptors. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT). IEEE **11**(6), 703-715 (2001)
186. Manjunath BS, Salembier P, Sikora T (2002) Introduction to MPEG-7: multimedia content description interface. John Wiley & Sons
187. Manning CD, Schütze H (1999) Foundations of statistical natural language processing. MIT press, Cambridge, MA, USA
188. Mao J, Xu W, Yang Y, Wang J, Yuille A L: Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090* (2014)
189. Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A: Deep captioning with multimodal recurrent neural networks (m-rnn). In: Proceedings of the 3rd International Conference on Learning Representations (ICLR). San Diego, CA, USA (2015)
190. Maree R, Geurts P, Piater J, Wehenkel L: Random subwindows for robust image classification. In: Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 34-40. IEEE, San Diego, CA, USA (2005)
191. Materka A, Strzelecki M: Texture analysis methods–a review. Technical university of lodz, institute of electronics, COST B11 report, Brussels, 9-11 (1998)
192. Mathews A P, Xie L, He X: SentiCap: Generating Image Descriptions with Sentiments. In: Proceedings of the 30th Association for the Advancement of Artificial Intelligence (AAAI), pp. 3574–3580. Phoenix, Arizona, USA (2016)
193. Mayhew M B, Chen B, Ni K S: Assessing semantic information in convolutional neural network representations of images via image annotation. In: Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), pp. 2266-2270. IEEE, Phoenix, AZ, USA (2016)
194. Mezaris V, Kompatsiaris I, Strintzis M G: An ontology approach to object-based image retrieval. In: Proceedings of the 2003 IEEE International Conference on Image Processing (ICIP), pp. 511-514. IEEE, Barcelona, Spain (2003)

195. Mezaris V, Kompatsiaris I, Strintzis MG (2004) Region-based image retrieval using an object ontology and relevance feedback. EURASIP Journal on Advances in Signal Processing Springer 2004(6):886–901
196. Mitran M, Mihalcea R, Cabanac G, Boughanem M: Landmark image annotation using textual and geolocation metadata. In: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval (OAIR), pp. 65-68. ACM, Lisbon, Portugal (2013)
197. Miyamori H, Iisaku S I: Video annotation for content-based retrieval using human behavior analysis and domain knowledge. In: Proceeding of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG), pp. 320-325. IEEE, Grenoble, France (2000)
198. Mori Y, Takahashi H, Oka R: Image-to-word transformation based on dividing and vector quantizing images with words. In: Proceedings of the 1st International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM), pp. 1-9. ACM, Orlando, Florida (1999)
199. Mousselly-Sergieh H, Egyed-Zsigmond E, Gianini G, Döller M, Kosch H, Pinon J M: Tag similarity in folksonomies. In: Proceedings of the XXXI INFORSID congress, pp. 319-334 (2013)
200. Muda Z, Lewis P H, Payne T R, Weal M J: Enhanced image annotations based on spatial information extraction and ontologies. In: Proceedings of the 2009 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), pp.173-178. IEEE, Kuala Lumpur, Malaysia (2009)
201. Murthy V N, Can E F, Manmatha R: A hybrid model for automatic image annotation. In: Proceedings of the 4th International Conference on Multimedia Retrieval (ICMR), pp. 369). ACM, Glasgow, UK (2014)
202. Murthy V N, Maji S, Manmatha R: Automatic image annotation using deep learning representations. In: Proceedings of the 5th ACM International Conference on Multimedia Retrieval (ICMR), pp. 603-606. ACM, Shanghai, China (2015)
203. Naik D., Shah P.: A review on image segmentation clustering algorithms. International Journal of Computer Science and Information Technologies (JCSIT). **5**(3), 3289-3289 (2014)
204. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E (2015) Deep learning applications and challenges in big data analytics. Journal of Big Data Springer 2(1):21
205. Nanda P. K, Ponacha P G, Desai U B: A Supervised Image Segmentation scheme using MRF Model and Homotopy Continuation Method. In: Proceedings of the *Indian Conference on Computer Vision*, *Graphics and Image Processing* (ICVGIP), pp. 15-20. Delhi, India (1998)
206. Natsev A, Rastogi R, Shim K: WALRUS: A similarity retrieval algorithm for image databases. In: Proceedings of the 1999 International Conference on Management of Data (ACM SIGMOD Record), pp. 395-406. ACM, Philadelphia, Pennsylvania, USA (1999)
207. Nguyen T V, Zhao Q, Yan S: Attentive systems: A survey. International Journal of Computer Vision (IJCV). **126**(1), 86-110 (2018)
208. Niles I, Pease A: Towards a standard upper ontology. In: Proceedings of the 2001 International Conference on Formal Ontology in Information Systems, pp. 2-9. ACM, Ogunquit, Maine, USA (2001)
209. Oberoi A, Singh M (2012) Content-based image retrieval system for medical data bases (CBIR-MD)-lucratively tested on endoscopy, dental and skull images. International Journal of Computer Science Issues (IJCSI) 9(3):300–306
210. Ojha U, Adhikari U, Singh D K: Image annotation using deep learning: A review. In: 2017 Proceedings of the International Conference on Intelligent Computing and Control (I2C2), pp. 1-5. IEEE, Coimbatore, India (2017)
211. Oliva D, Cuevas E: An Introduction to Machine Learning. Advances and Applications of Optimized Algorithms in Image Processing, pp.1–11. Springer Vol. 117 (2017)
212. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175
213. Ordonez V, Kulkarni G, Berg T L: Im2text: Describing images using 1 million captioned photographs. In: Proceedings of the 25th Advances in Neural Information Processing Systems (NIPS), pp. 1143-1151. Granada, Spain (2011)
214. Panda S: Unsupervised Color Image Segmentation using MRF Models to Preserve Weak Edges. International Journal of Computer & Mathematical Sciences (IJCMS). **5**(6), 73-81 (2016)
215. Pandey S, Khanna P: A hierarchical clustering approach for image datasets. In: Proceedings of the 9th International Conference on Industrial and Information Systems (ICIIS), pp. 1-6. IEEE, Gwalior, India (2014)
216. Park SB, Lee JW, Kim SK (2004) Content-based image classification using a neural network. Pattern Recognition Letters Elsevier science 25(3):287–300
217. Pass G, Zabih R: Histogram refinement for content-based image retrieval. In: Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV), pp. 96-102. IEEE, Sarasota, FL, USA (1996)
218. Pass G, Zabih R (1999) Comparing images using joint histograms. Multimedia systems Springer 7(3):234–240

219. Patil MP, Kolhe SR (2012) Automatic image categorization and annotation using K-NN for COREL dataset. Advances in Computational Research 4(1):108–112
220. Patil M P, Kolhe S R: Automatic Image Annotation Using Decision Trees and Rough Sets. International Journal of Computer Science & Applications (IJCSA). **11**(2), 38-49 (2014)
221. Pawlak Z (1982) Rough sets. International Journal of Computer & Information Sciences Springer 11(5): 341–356
222. Peleg S, Naor J, Hartley R, Avnir D: Multiple resolution texture analysis and classification. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). IEEE **6**(4), 518-523 (1984)
223. Perronnin F, Sánchez J, Mensink T: Improving the fisher kernel for large-scale image classification. In: Proceedings of the 11th European Conference on Computer Vision (ECCV), pp. 143-156. Crete, Greece (2010)
224. Petridis K, Anastasopoulos D, Saathoff C, Timmermann N, Kompatsiaris Y, Staab S: M-OntoMat-Annotizer: Image annotation linking ontologies and multimedia low-level features. In: Proceedings of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES), pp. 633-640. Springer, Bournemouth, UK (2006)
225. Ping Tian D: A review on image feature extraction and representation techniques. International Journal of Multimedia and Ubiquitous Engineering (*IJMUE*). **8**(4), 385-396 (2013)
226. Pinheiro, P. O., Collobert, R., Dollár, P.: Learning to segment object candidates. In: Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), pp (1990-1998) IEEE, Montreal. Canada 2015
227. Preece J, Elser J, Jaiswal P, Kvilekval K, Fedorov D, Manjunath BS, Kitchen R, Xu X, Trigkakis D, Todorovic S, Carbon S (2016) Plant image segmentation and annotation with ontologies in BisQue. In: proceedings of the 7th joint international conference on biological ontology and BioCreative (ICBO/BioCreative). Corvallis. Oregon
228. Qi X, Han Y (2007) Incorporating multiple SVMs for automatic image annotation. Pattern Recognition Elsevier science 40(2):728–741
229. Qian Y, Zhou W, Yan J, Li W, Han L (2015) Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. Remote sensing of Environment Elsevier science 7(1):153–168
230. Qiu B: A refined SVM applied in medical image annotation. In: Proceedings of the Workshop of the Cross-Language Evaluation Forum for European Languages, pp. 690-693. Springer, Alicante, Spain (2006)
231. Quattrone G, Ferrara E, De Meo P, Capra L: Measuring similarity in large-scale folksonomies. In: Proceedings of the 23rd International Conference on Software Engineering and Knowledge Engineering (SEKE), pp. 385-391. Miami Beach, USA (2012)
232. Quinlan JR (1986) Induction of decision trees. Machine learning Springer 1(1):81–106
233. Quinlan J R: C4.5: Programs for Machine Learning, Morgan Kaufmann, Los Altos, California, USA (1993)
234. Redmon J, Farhadi A: YOLO9000: better, faster, stronger. In: Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7263-7271. IEEE, Honolulu, HI, USA (2017)
235. Redmon J, Divvala S, Girshick R, Farhadi A: You only look once: Unified, real-time object detection. In: Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788. IEEE, Las Vegas, NV, USA (2016)
236. Ren S, He K, Girshick R, Sun J: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proceedings of the 28th Advances in Neural Information Processing Systems (NIPS), pp. 91-99. Montreal, Quebec, Canada (2015)
237. Ren Z, Wang X, Zhang N, Lv X, Li L J: Deep reinforcement learning-based image captioning with embedding reward. In: Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 290-298. IEEE, Honolulu, HI, USA (2017)
238. Rennie S J, Marcheret E, Mroueh Y, Ross J, Goel V: Self-critical sequence training for image captioning. In: Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1179–1195. IEEE, Honolulu, HI, USA (2017)
239. Rosenfeld A, Weszka J S: Picture recognition. Digital Pattern Recognition. Springer, p. 135-166 (1980)
240. Rubner, Y., Tomasi, C., Guibas, L. J.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision (*IJCV*). Springer **40**(2), 99-121 (2000)
241. Rui Y, Huang T S, Ortega M, Mehrotra S: Relevance feedback: a power tool for interactive content-based image retrieval. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT). IEEE **8**(5), 644-655 (1998)
242. Rui Y, Huang T S, Chang S F: Image retrieval: Current techniques, promising directions, and open issues. Journal of Visual Communication and Image Representation (JVCI). Elsevier science **10**(1), 39-62 (1999)

243. Rui S, Jin W, Chua T S: A novel approach to auto image annotation based on pairwise constrained clustering and semi-naïve Bayesian model. In: Proceedings of the 11th International Conference on Multimedia Modelling (MMM), pp. 322–327. IEEE, Melbourne, Australia (2005)
244. Russell BC, Torralba A, Murphy KP, Freeman WT (2008) LabelMe: a dadabase and web-based tool or image annotation. International Journal of Computer Vision Springer 77(1-3):157–173
245. Sak H, Senior A, Beaufays F: Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. CoRR, abs/1402.1128 (2014)
246. Sami M, El-Bendary N, Hassanien A E: Automatic image annotation via incorporating Naive Bayes with particle swarm optimization. In: Proceedings of the World Congress on Information and Communication Technologies (WICT), pp. 790-794. IEEE, Trivandrum, India (2012)
247. Senthilkumar R, Prakash T S: Image Retrieval System by Automatic Annotation. International Journal on Engineering Technology and Sciences (IJETS). **1**(8), 286-290 (2014)
248. Senthilkumaran N, Vaithegi S: Image segmentation by using thresholding techniques for medical images. International Journal of Computer Science and Engineering (IJCSE). **6**(1), 1-13 (2016)
249. Serrano N, Savakis A, Luo A: A computationally efficient approach to indoor/outdoor scene classification. In: Proceedings of the 16th International Conference on Pattern Recognition (ICPR), pp. 146-149. IEEE, Quebec City, Quebec, Canada (2002)
250. Sethi I K, Coman I L, Stan D: Mining association rules between low-level image features and high-level concepts. In: International Society for Optics and Photonics (SPIE). Vol. 4384, pp. 279-291 (2001)
251. Shen J, Wang M, Yan S, Hua X S: Multimedia tagging: past, present and future. In: Proceedings of the 19th ACM International Conference on Multimedia (MM), pp. 639-640. ACM, Scottsdale, AZ, USA (2011)
252. Shen Z, Liu Z, Li J, Jiang Y G, Chen Y, Xue X. Dsod: Learning deeply supervised object detectors from scratch. In: Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), pp. 1919-1927. IEEE, Venice, Italy (2017)
253. Shetty R, Rohrbach M, Anne Hendricks L, Fritz M, Schiele B.: Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training. In: Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), pp. 4155–4164. IEEE, Venice, Italy (2017)
254. Shi J, Malik J: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). IEEE **22**(8), 888-905 (2000)
255. Shi R, Feng H, Chua T S, Lee C H: An adaptive image content representation and segmentation approach to automatic image annotation. In: Proceedings of the 3rd International Conference on Image and Video Retrieval (CIVR), pp. 545-554. Springer, Dublin, Ireland (2004)
256. Shimpi S, Patil V: Hidden Markov model as classifier: a survey. In: Proceedings of the 2013 International Conference on Computer Science and Engineering (COMPSE), pp. 13530-13533 (2013)
257. Shitole A, Godase U: Survey on Content Based Image Retrieval. International Journal of Computer-Aided Technologies (IJCAx). **1**(1), 21-29 (2014)
258. Shukla T, Mishra N, Sharma S (2013) Automatic image annotation using SURF features. Int J Comput Appl 68(4):17–24
259. Shyu C R: Relevance feedback decision trees in content-based image retrieval. In: Proceedings of the 2000 IEEE Workshop on Content-based Access of Image and Video Libraries, pp. 68-72. IEEE, Hilton Head Island, SC, USA (2000)
260. Simonyan K, Zisserman A: Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556 (2014)
261. Smeulders AW, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence IEEE 22(12):1349–1380
262. Socher R, Perelygin A, Wu J, Chuang J, Manning C D, Ng A, Potts C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1631-1642. Washington, USA (2013)
263. Sreedhar Kumar S, Shilpa S.: A new approach for image feature vector classification using unsupervised clustering method. International Journal of Advance Research in Science And Engineering (IJARSE). **3**(6), 108-117 (2014)
264. Stanchev PL, Green D Jr, Dimitrov B (2003) Level color similarity retrieval. International Journal of Information Theories & Application 10(3):363–369
265. Steggink J, Snoek CG (2011) Adding semantics to image-region annotations with the name-it-game. Multimedia Systems Springer 17(5):367–378
266. Stührenberg M (2013) What, when, where? Spatial and temporal annotations with XStandoff. In Balisage, The Markup Conference. Montréal, Canada
267. Sugano Y, Bulling A: Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv: 1608.05203* (2016)

268. Sun C, Gan C, Nevatia R.: Automatic concept discovery from parallel text and visual corpora. In: Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV), pp. 2596–2604. IEEE, Santiago, Chile (2015)
269. Swain M J, Ballard D H: Color indexing. International Journal of Computer Vision (*IJCV*). Springer **7**(1), 11-32 (1991)
270. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Rabinovich A: Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-9. IEEE, Boston, MA, USA (2015)
271. Tabb M, Ahuja N: Multiscale image segmentation by integrated edge and region detection. IEEE Transactions on Image Processing (TIP). IEEE **6**(5), 642-655 (1997)
272. Tallapragada V S, Reddy D M, Kiran P S, Reddy D V: A Novel Medical Image Segmentation and Classification using Combined Feature Set and Decision Tree Classifier. International Journal of Research in Engineering and Technology (*IJRET*). **4**(9), 83-86 (2016)
273. Tamura H, Mori S, Yamawaki T: Textural features corresponding to visual perception. IEEE Transactions on Systems, Man, and Cybernetics. IEEE **8**(6), 460-473 (1978)
274. Tan, W., Wang, X., Zhang, Y., Zhou, B., Chen, X.: A conceptual prototype for digital media cloud. In: Proceedings of the 8th ChinaGrid Annual Conference (ChinaGrid), pp. 103-108. IEEE, Changchun, China (2013)
275. Tang J, Hong R, Yan S, Chua TS, Qi GJ, Jain R (2011) Image annotation by k nn-sparse graph-based label propagation over noisily tagged web images. ACM Transactions on Intelligent Systems and Technology (TIST) 2(2):1–15
276. Tang J, Chen Q, Wang M, Yan S, Chua TS, Jain R (2013) Towards optimizing human labeling for interactive image tagging. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 9(4):1–18
277. Tang J, Yan S, Zhao C, Chua TS, Jain R (2013) Label-specific training set construction from web resource for image annotation. Signal Processing (SP) 93(8):2199–2204
278. Tian D: Support vector machine for automatic image annotation. International Journal of Hybrid Information Technology (*IJHIT). ***8**(11), 435-446 (2015)
279. Tian Z, Shen C, Chen H, He T.: FCOS: Fully Convolutional One-Stage Object Detection. *arXiv preprint arXiv:1904.01355* (2019)
280. Ting Y, Yingwei P, Yehao L, Zhaofan Q, and Tao M: Boosting image captioning with attributes. In: Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), pp. 4904–4912. IEEE, Venice, Italy (2017)
281. Torralba A, Russell BC, Yuen J (2010) Labelme: online image annotation and applications. Proc IEEE 98(8):1467–1484
282. Town C, Sinclair D (2000) Content based image retrieval using semantic visual categories. Society of Manufacturing Engineers
283. Tran K, He X, Zhang L, Sun J: Rich image captioning in the wild. In: Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 49–56. IEEE, Las Vegas, NV, USA (2016)
284. Trelea IC (2003) The particle swarm optimization algorithm: convergence analysis and parameter selection. Information processing letters Elsevier science 85(6):317–325
285. Tsai C F, McGarry K, Tait J: CLAIRE: A modular support vector image indexing and classification system. ACM Transactions on Information Systems (TOIS). ACM **24**(3), 353-379 (2006)
286. Tuceryan M, Jain A K: Texture analysis. In: Handbook of Pattern Recognition and Computer Vision, pp. 235-276 (1993)
287. Tunga S, Jayadevappa D, Gururaj C: A comparative study of content based image retrieval trends and approaches. International Journal of Image Processing (IJIP). **9**(3), 127-155 (2015)
288. Tyagi V: Content-Based Image Retrieval Techniques: A Review. In: Proceeding of the 2017 Content-Based Image Retrieval, pp. 29-48. Springer, Singapore (2017)
289. Ugarriza L G, Saber E, Vantaram S R, Amuso V, Shaw M, Bhaskar R: Automatic image segmentation by dynamic region growth and multiresolution merging. IEEE Transactions on Image Processing (TIP). IEEE **18**(10), 2275-2288 (2009)
290. Uijlings JR, Van De Sande KE, Gevers T, Smeulders AW (2013) Selective search for object recognition. International Journal of Computer Vision (IJCV) 104(2):154–171
291. Vedaldi A, Gulshan V, Varma M, Zisserman A: Multiple kernels for object detection. In: Proceedings of the 12th IEEE International Conference on Computer Vision (ICCV), pp. 606-613. IEEE, Kyoto, Japan (2009)
292. Vega F, Pérez W, Tello A, Saquicela V, Espinoza M, Vidal M, La Cruzc A: WebMedSA: a web-based framework for segmenting and annotating medical images using biomedical ontologies. In: Proceedings of

the 11th International Symposium on Medical Information Processing and Analysis (SIPAIM), pp. 134-146, Cuenca, Ecuador (2015)

293. Venugopalan S, Hendricks L A, Rohrbach M, Mooney R, Darrell T, Saenko K: Captioning images with diverse objects. In: Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1170–1178. IEEE, Honolulu, HI, USA (2017)
294. Verma Y, Jawahar C V: Image annotation using metric learning in semantic neighbourhoods. In: Proceedings of the 12th European Conference on Computer Vision (ECCV), pp. 836-849. Springer, Firenze, Italy (2012)
295. Vincent L, Soille P: Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI). IEEE **13**(6), 583-598 (1991)
296. Visa A, Valkealahti K, Simula O: Cloud detection based on texture segmentation by neural network methods. In: Proceedings of the 1991 IEEE International Conference Joint Conference on Neural Networks (IJCNN), pp. 1001-1006. IEEE, Singapore (1991)
297. Von Ahn L, Dabbish L: Labeling images with a computer game. In: Proceedings of the 2004 ACM Conference on Human Factors in Computing Systems, pp. 319-326. ACM, Vienna, Austria (2004)
298. Von Ahn L, Liu R, Blum M: Peekaboom: A game for locating objects in images. In: Proceedings of the 2006 ACM SIGCHI conference on Human in Computing Systems, pp. 55–64. ACM, Montréal, Québec, Canada (2006)
299. Wagstaff K, Cardie C, Rogers S, Schrödl S: Constrained K-means Clustering with Background Knowledge. In: Proceedings of the 18th International Conference on Machine Learning (ICML), pp. 577-584. ACM, Williamstown, MA, USA (2001)
300. Wang Q, Chan A B: CNN+ CNN: convolutional decoders for image captioning. *arXiv preprint arXiv: 1805.09019* (2018)
301. Wang J Z, Li J: Learning-based linguistic indexing of pictures with 2–d MHMMs. In: Proceedings of the 10th ACM International Conference on Multimedia (MM), pp. 436-445. ACM, Juan-les-Pins, France (2002)
302. Wang J Z, Li J, Wiederhold G: SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). IEEE **23**(9), 947-963 (2001)
303. Wang C, Yan S, Zhang L, Zhang H J: Multi-label sparse coding for automatic image annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1643-1650. IEEE, Miami, FL, USA (2009)
304. Wang T, Wu D J, Coates A, Ng A Y: End-to-end text recognition with convolutional neural networks. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR), pp. 3304-3308. IEEE, Tsukuba, Japan (2012)
305. Wang XY, Zhang BB, Yang HY (2014) Content-based image retrieval by integrating color and texture features. Multimedia Tools and Applications Springer 68(3):545–569
306. Wang R, Xie Y, Yang J, Xue L, Hu M, Zhang Q: Large scale automatic image annotation based on convolutional neural network. Journal of Visual Communication and Image Representation (JVCI). Elsevier science **49**(C), 213-224 (2017)
307. Wei Z, Luo X, Zhou F: Ontology based automatic image annotation using multi-class SVM. In: Proceedings of the 7th International Conference on Image and Graphics (ICIG), pp. 434-438. IEEE, Qingdao, China (2013)
308. Wei Y, Liang X, Chen Y, Jie Z, Xiao Y, Zhao Y, Yan S (2016) Learning to segment with image-level annotations. Pattern Recognition (PR) 59:234–244
309. Wei C, Huang J, Mansaray LR, Li Z, Liu W, Han J (2017) Estimation and mapping of winter oilseed rape LAI from high spatial resolution satellite data based on a hybrid method. Remote sensing of Environment Elsevier science 9(5):488
310. Wei-ning W, Ying-lin Y, Sheng-ming J: Image retrieval by emotional semantics: A study of emotional space and feature extraction. In: Proceedings of the 2006 IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 3534-3539. IEEE, Taipei, Taiwan (2006)
311. Weston J, Bengio S, Usunier N: Wsabie: Scaling up to large vocabulary image annotation. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI), pp. 2764-2770. ACM, Barcelona, Catalonia, Spain (2011)
312. Wojnar A, Pinheiro A M: Annotation of medical images using the SURF descriptor. In: Proceedings of the 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 130-133. IEEE, Barcelona, Spain (2012)
313. Wong R C, Leung C H: Automatic semantic annotation of real-world web images. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). IEEE 30(11), 1933-1944 (2008)
314. Wong ST, Tjandra DA (1999) A digital library for biomedical imaging on the internet. IEEE Commun Mag 37(1):84–91

315. Wu J, Yu Y, Huang C, Yu K: Deep multiple instance learning for image classification and auto-annotation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3460-3469. IEEE, Boston, MA, USA (2015)
316. Xu H, Zhou X, Wang M, Xiang Y, Shi B: Exploring Flickr's related tags for semantic annotation of web images. In: Proceedings of the 2009 ACM International Conference on Image and Video Retrieval (CIVR), p. 46. ACM, Santorini, Fira, Greece (2009)
317. Xu Z, Luo X, Liu Y, Mei L, Hu C (2014) Measuring semantic relatedness between flickr images: from a social tag based view. Sci World J 2014(758089)
318. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Bengio Y: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning (ICML), pp. 2048–2057. Lille, France (2015)
319. Xue J, Li J, Gong Y.: Restructuring of deep neural network acoustic models with singular value decomposition. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech), pp. 2365-2369. Lyon, France (2013)
320. Yang C, Dong M, Fotouhi F: Image content annotation using bayesian framework and complement components analysis. In: Proceedings of the 2005 IEEE International Conference on Image Processing (ICIP), pp. pp. 1190-1193. IEEE, Genova, Italy (2005)
321. Yang C, Dong M, Hua J.: Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2057-2063. IEEE, New York, NY, USA (2006)
322. Yang M, Kpalma K, Ronsin J: A survey of shape feature extraction techniques. Pattern Recognition. Elsevier science p. 43-90. (2008).
323. Yang Y, Zhang W, Xie Y (2015) Image automatic annotation via multi-view deep representation. Journal of Visual Communication and Image Representation Elsevier science/ACM 33(2015):368–377
324. Yang L, Tang K, Yang J, Li L J.: Dense Captioning with Joint Inference and Visual Context. In: Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1978-1987. IEEE, Honolulu, HI, USA (2017)
325. Yao T, Pan Y, Li Y, Mei T: Incorporating copying mechanism in image captioning for learning novel objects. In: Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5263–5271. IEEE, Honolulu, HI, USA (2017)
326. Yavlinsky A, Schofield E, Rüger S: Automated image annotation using global features and robust nonparametric density estimation. In: Proceedings of the 4th International Conference on Image and Video Retrieval (CIVR), pp. 507-517. Springer, Singapore (2005)
327. You, D., Antani, S., Demner-Fushman, D., Thoma, G. R.: A contour-based shape descriptor for biomedical image classification and retrieval. Document Recognition and Retrieval (DRR). Vol. 9021, p. 90210L (2014)
328. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4651-4659. IEEE, Las Vegas, NV, USA (2016)
329. Yue J, Li Z, Liu L, Fu Z (2011) Content-based image retrieval using color and texture fused features. Mathematical and Computer Modelling Elsevier science 54(3-4):1121–1127
330. Zahn C T: Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on Computers (TC). IEEE **20**(1), 68–86 (1971)
331. Zhang H: The Optimality of Naive Bayes. In: Proceedings of the 17th International Conference of Florida AI Research Society (FLAIRS), pp. 17-19. Florida, USA (2004)
332. Zhang D, Lu G (2004) Review of shape representation and description techniques. Pattern recognition Elsevier science 37(1):1–19
333. Zhang ML, Zhou ZH (2007) ML-KNN: a lazy learning approach to multi-label learning. Pattern Recognition Elsevier science 40(7):2038–2048
334. Zhang T, Ramakrishnan R, Livny M (1996) BIRCH: an efficient data clustering method for very large databases. In: ACM Sigmod Record ACM 25(2):103–114
335. Zhang C, Chai J, Jin R: User term feedback in interactive text-based image retrieval. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 51-58. ACM, Salvador, Brazil (2005)
336. Zhang D, Islam MM, Lu G (2012) A review on automatic image annotation techniques. Pattern Recognition Elsevier science 45(1):346–362
337. Zhao Y, Zhao Y, Zhu Z (2009) TSVM-HMM: Transductive SVM based hidden Markov model for automatic image annotation. Expert Systems with Applications Elsevier science 36(6):9813–9818