UNIT-1V

NATURAL LANGUAGE FOR COMMUNICATION

Concept of Grammar

Grammar is very essential and important to describe the syntactic structure of well-formed programs. In the literary sense, they denote syntactical rules for conversation in natural languages. Linguistics have attempted to define grammars since the inception of natural languages like English, Hindi, etc.

The theory of formal languages is also applicable in the fields of Computer Science mainly in programming languages and data structure. For example, in ‘C’ language, the precise grammar rules state how functions are made from lists and statements.

A mathematical model of grammar was given by **Noam Chomsky** in 1956, which is effective for writing computer languages.

Mathematically, a grammar G can be formally written as a 4-tuple (N, T, S, P) where −

- **N** or $\mathbf{V_N}$ = set of non-terminal symbols, i.e., variables.
- **T** or ∑ = set of terminal symbols.
- **S** = Start symbol where S ∈ N
- **P** denotes the Production rules for Terminals as well as Non-terminals. It has the form α → β, where α and β are strings on V_N ∪ ∑ and least one symbol of α belongs to V_N

Phrase Structure or Constituency Grammar

Phrase structure grammar, introduced by Noam Chomsky, is based on the constituency relation. That is why it is also called constituency grammar. It is opposite to dependency grammar.

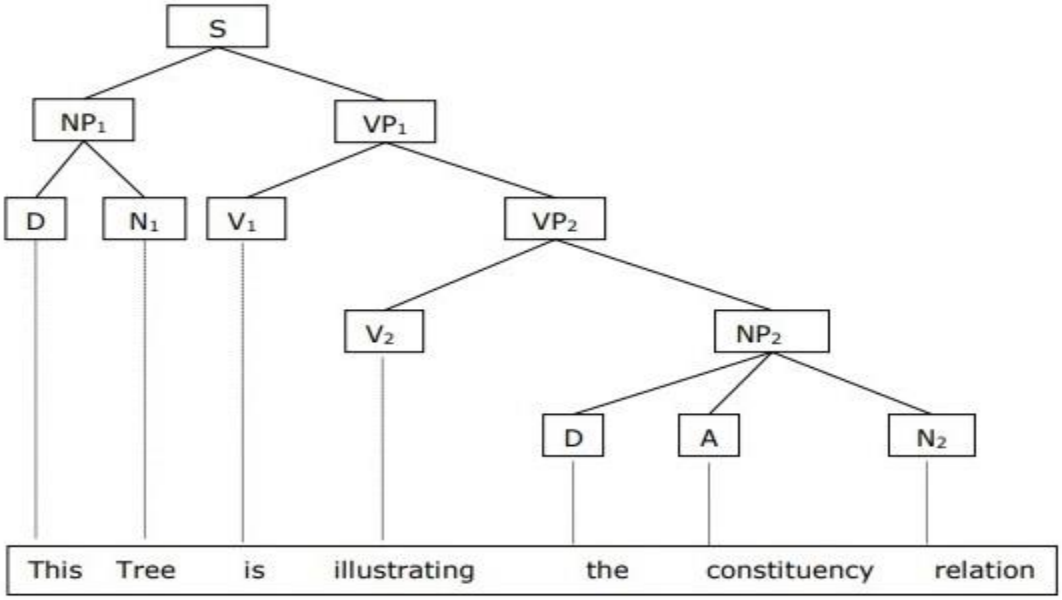
Example

Before giving an example of constituency grammar, we need to know the fundamental points about constituency grammar and constituency relation.

- All the related frameworks view the sentence structure in terms of constituency relation.
- The constituency relation is derived from the subject-predicate division of Latin as well as Greek grammar.

- The basic clause structure is understood in terms of **noun phrase NP** and **verb phrase VP**.

We can write the sentence **"This tree is illustrating the constituency relation"** as follows −



Dependency Grammar

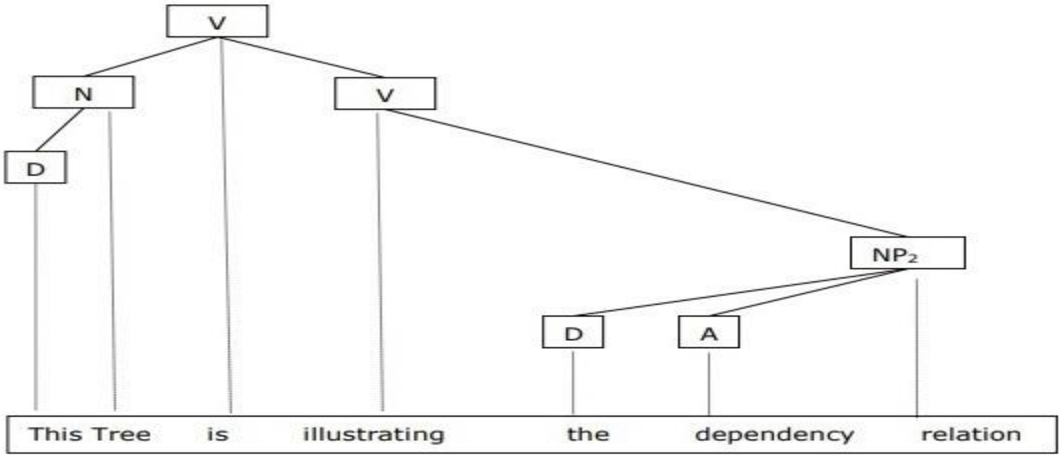
It is opposite to the constituency grammar and based on dependency relation. It was introduced by Lucien Tesniere. Dependency grammar (DG) is opposite to the constituency grammar because it lacks phrasal nodes.

Example

Before giving an example of Dependency grammar, we need to know the fundamental points about Dependency grammar and Dependency relation.

- In DG, the linguistic units, i.e., words are connected to each other by directed links.
- The verb becomes the center of the clause structure.
- Every other syntactic units are connected to the verb in terms of directed link. These syntactic units are called **dependencies**.

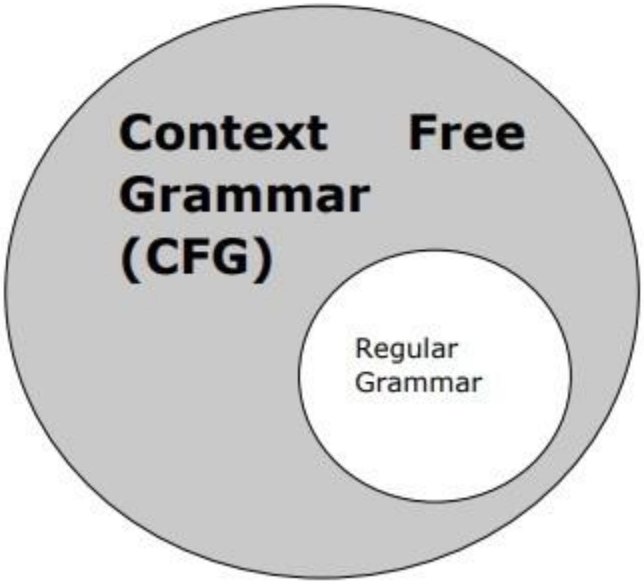
We can write the sentence **"This tree is illustrating the dependency relation"** as follows;



Parse tree that uses Constituency grammar is called constituency-based parse tree; and the parse trees that uses dependency grammar is called dependency-based parse tree.

Context Free Grammar

Context free grammar, also called CFG, is a notation for describing languages and a superset of Regular grammar. It can be seen in the following diagram −



Definition of CFG

CFG consists of finite set of grammar rules with the following four components −

Set of Non-terminals

It is denoted by V. The non-terminals are syntactic variables that denote the sets of strings, which further help defining the language, generated by the grammar.

Set of Terminals

It is also called tokens and defined by Σ. Strings are formed with the basic symbols of terminals.

Set of Productions

It is denoted by P. The set defines how the terminals and non-terminals can be combined. Every production(P) consists of non-terminals, an arrow, and terminals (the sequence of terminals). Non-terminals are called the left side of the production and terminals are called the right side of the production.

Start Symbol

The production begins from the start symbol. It is denoted by symbol S. Non-terminal symbol is always designated as start symbol.

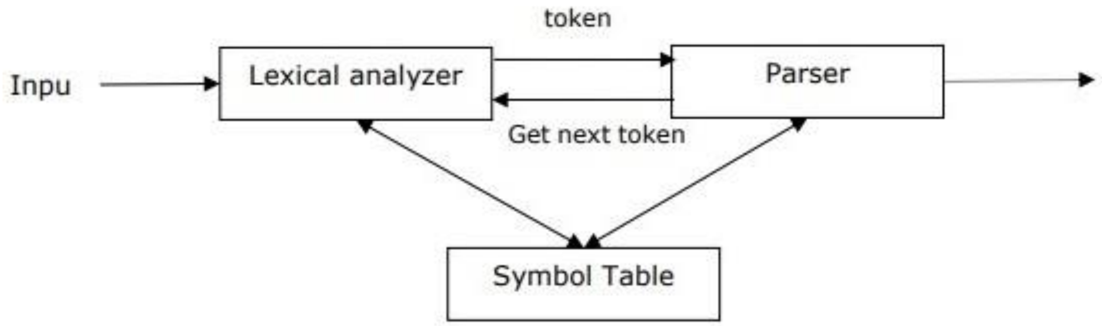
Syntactic Analysis

Syntactic analysis or parsing or syntax analysis is the third phase of NLP. The purpose of this phase is to draw exact meaning, or you can say dictionary meaning from the text. Syntax analysis checks the text for meaningfulness comparing to the rules of formal grammar. For example, the sentence like “hot ice-cream” would be rejected by semantic analyzer.

In this sense, syntactic analysis or parsing may be defined as the process of analyzing the strings of symbols in natural language conforming to the rules of formal grammar. The origin of the word **‘parsing’** is from Latin word **‘pars’** which means **‘part’**.

Concept of Parser

It is used to implement the task of parsing. It may be defined as the software component designed for taking input data (text) and giving structural representation of the input after checking for correct syntax as per formal grammar. It also builds a data structure generally in the form of parse tree or abstract syntax tree or other hierarchical structure.



The main roles of the parse include −

- To report any syntax error.
- To recover from commonly occurring error so that the processing of the remainder of program can be continued.
- To create parse tree.
- To create symbol table.
- To produce intermediate representations (IR).

Types of Parsing

Derivation divides parsing into the followings two types −

- Top-down Parsing
- Bottom-up Parsing

Top-down Parsing

In this kind of parsing, the parser starts constructing the parse tree from the start symbol and then tries to transform the start symbol to the input. The most common form of topdown parsing uses recursive procedure to process the input. The main disadvantage of recursive descent parsing is backtracking.

Bottom-up Parsing

In this kind of parsing, the parser starts with the input symbol and tries to construct the parser tree up to the start symbol.

Concept of Derivation

In order to get the input string, we need a sequence of production rules. Derivation is a set of production rules. During parsing, we need to decide the non-terminal, which is to be replaced along with deciding the production rule with the help of which the non-terminal will be replaced.

Types of Derivation

In this section, we will learn about the two types of derivations, which can be used to decide which non-terminal to be replaced with production rule −

Left-most Derivation

In the left-most derivation, the sentential form of an input is scanned and replaced from the left to the right. The sentential form in this case is called the left-sentential form.

Right-most Derivation

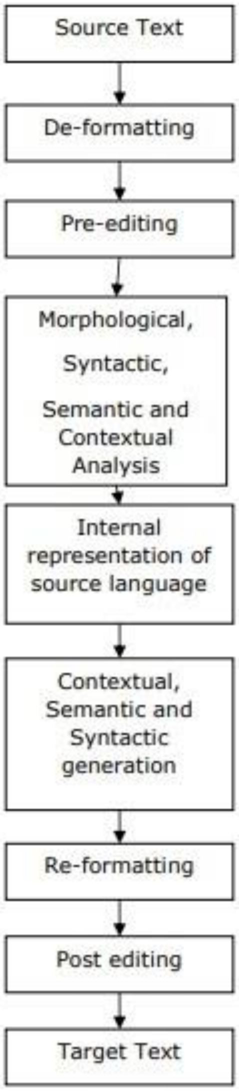
In the left-most derivation, the sentential form of an input is scanned and replaced from right to left. The sentential form in this case is called the right-sentential form.

Concept of Parse Tree

It may be defined as the graphical depiction of a derivation. The start symbol of derivation serves as the root of the parse tree. In every parse tree, the leaf nodes are terminals and interior nodes are non-terminals. A property of parse tree is that in-order traversal will produce the original input string.

MACHINE TRANSLATION

Machine translation (MT), process of translating one source language or text into another language, is one of the most important applications of NLP. We can understand the process of machine translation with the help of the following flowchart −


Source Text
De-formatting
Pre-editing
Morphological,
Syntactic,
Semantic and Contextual Analysis
Internal representation of source language
Contextual, Semantic and Syntactic generation
Re-formatting
Post editing
Target Text

Types of Machine Translation Systems

There are different types of machine translation systems. Let us see what the different types are.

Bilingual MT System

Bilingual MT systems produce translations between two particular languages.

Multilingual MT System

Multilingual MT systems produce translations between any pair of languages. They may be either uni-directional or bi-directional in nature.

Approaches to Machine Translation (MT)

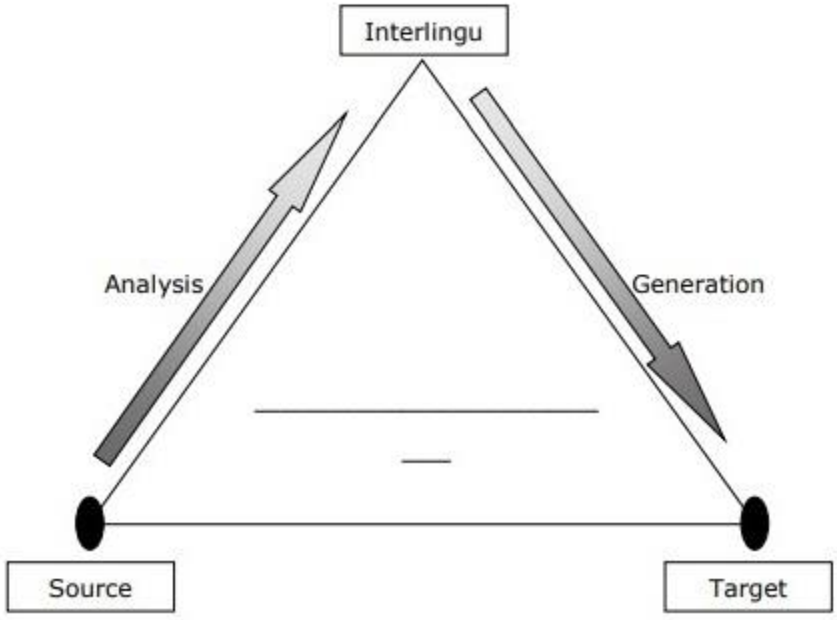
Let us now learn about the important approaches to Machine Translation. The approaches to MT are as follows −

Direct MT Approach

It is less popular but the oldest approach of MT. The systems that use this approach are capable of translating SL (source language) directly to TL (target language). Such systems are bi-lingual and uni-directional in nature.

Interlingua Approach

The systems that use Interlingua approach translate SL to an intermediate language called Interlingua (IL) and then translate IL to TL. The Interlingua approach can be understood with the help of the following MT pyramid −



Transfer Approach

Three stages are involved with this approach.

- In the first stage, source language (SL) texts are converted to abstract SL-oriented representations.

- In the second stage, SL-oriented representations are converted into equivalent target language (TL)-oriented representations.
- In the third stage, the final text is generated.

Empirical MT Approach

This is an emerging approach for MT. Basically, it uses large amount of raw data in the form of parallel corpora. The raw data consists of the text and their translations. Analogybased, example-based, memory-based machine translation techniques use empirical MTapproach.

Speech recognition

Speech recognition, also known as automatic speech recognition (ASR), computer speech recognition, or speech-to-text, is a capability which enables a program to process human speech into a written format. While it's commonly confused with voice recognition, speech recognition focuses on the translation of speech from a verbal format to a text one whereas voice recognition just seeks to identify an individual user's voice.

Key features of effective speech recognition

Many speech recognition applications and devices are available, but the more advanced solutions use AI and machine learning. They integrate grammar, syntax, structure, and composition of audio and voice signals to understand and process human speech. Ideally, they learn as they go — evolving responses with each interaction.

The best kind of systems also allow organizations to customize and adapt the technology to their specific requirements — everything from language and nuances of speech to brand recognition. For example:

- **Language weighting:** Improve precision by weighting specific words that are spoken frequently (such as product names or industry jargon), beyond terms already in the base vocabulary.
- **Speaker labeling:** Output a transcription that cites or tags each speaker's contributions to a multi-participant conversation.
- **Acoustics training:** Attend to the acoustical side of the business. Train the system to adapt to an acoustic environment (like the ambient noise in a call center) and speaker styles (like voice pitch, volume and pace).
- **Profanity filtering:** Use filters to identify certain words or phrases and sanitize speech output.

Meanwhile, speech recognition continues to advance. Companies, like IBM, are making inroads in several areas, the better to improve human and machine interaction.

Speech recognition algorithms

The vagaries of human speech have made development challenging. It's considered to be one of the most complex areas of computer science – involving linguistics, mathematics and statistics. Speech recognizers are made up of a few components, such as the speech input, feature extraction, feature vectors, a decoder, and a word output. The decoder leverages acoustic models, a pronunciation dictionary, and language models to determine the appropriate output.

Speech recognition technology is evaluated on its accuracy rate, i.e. word error rate (WER), and speed. A number of factors can impact word error rate, such as pronunciation, accent, pitch, volume, and background noise. Reaching human parity – meaning an error rate on par with that of two humans speaking – has long been the goal of speech recognition systems. Research from Lippmann (link resides outside IBM) (PDF, 344 KB) estimates the word error rate to be around 4 percent, but it's been difficult to replicate the results from this paper.

Various algorithms and computation techniques are used to recognize speech into text and improve the accuracy of transcription. Below are brief explanations of some of the most commonly used methods:

- **Natural language processing (NLP):** While NLP isn't necessarily a specific algorithm used in speech recognition, it is the area of artificial intelligence which focuses on the interaction between humans and machines through language through speech and text. Many mobile devices incorporate speech recognition into their systems to conduct voice search—e.g. Siri—or provide more accessibility around texting.
- **Hidden markov models (HMM):** Hidden Markov Models build on the Markov chain model, which stipulates that the probability of a given state hinges on the current state, not its prior states. While a Markov chain model is useful for observable events, such as text inputs, hidden markov models allow us to incorporate hidden events, such as part-of-speech tags, into a probabilistic model. They are utilized as sequence models within speech recognition, assigning labels to each unit—i.e. words, syllables, sentences, etc.—in the sequence. These labels create a mapping with the provided input, allowing it to determine the most appropriate label sequence.
- **N-grams:** This is the simplest type of language model (LM), which assigns probabilities to sentences or phrases. An N-gram is sequence of N-words. For example, "order the pizza" is a trigram or 3-gram and "please order the pizza" is a 4-gram. Grammar and the probability of certain word sequences are used to improve recognition and accuracy.

- **Neural networks:** Primarily leveraged for deep learning algorithms, neural networks process training data by mimicking the interconnectivity of the human brain through layers of nodes. Each node is made up of inputs, weights, a bias (or threshold) and an output. If that output value exceeds a given threshold, it "fires" or activates the node, passing data to the next layer in the network. Neural networks learn this mapping function through supervised learning, adjusting based on the loss function through the process of gradient descent. While neural networks tend to be more accurate and can accept more data, this comes at a performance efficiency cost as they tend to be slower to train compared to traditional language models.
- **Speaker Diarization (SD):** Speaker diarization algorithms identify and segment speech by speaker identity. This helps programs better distinguish individuals in a conversation and is frequently applied at call centers distinguishing customers and sales agents.

SPEECH RECOGNITION USE CASES

- A wide number of industries are utilizing different applications of speech technology today, helping businesses and consumers save time and even lives. Some examples include:
- **Automotive:** Speech recognizers improves driver safety by enabling voice-activated navigation systems and search capabilities in car radios.
- **Technology:** Virtual agents are increasingly becoming integrated within our daily lives, particularly on our mobile devices. We use voice commands to access them through our smartphones, such as through Google Assistant or Apple's Siri, for tasks, such as voice search, or through our speakers, via Amazon's Alexa or Microsoft's Cortana, to play music. They'll only continue to integrate into the everyday products that we use, fueling the "Internet of Things" movement.
- **Healthcare:** Doctors and nurses leverage dictation applications to capture and log patient diagnoses and treatment notes.
- **Sales:** Speech recognition technology has a couple of applications in sales. It can help a call center transcribe thousands of phone calls between customers and agents to identify common call patterns and issues. AI chatbots can also talk to people via a webpage, answering common queries and solving basic requests without needing to wait for a contact center agent to be available. It both instances speech recognition systems help reduce time to resolution for consumer issues.
- **Security:** As technology integrates into our daily lives, security protocols are an increasing priority. Voice-based authentication adds a viable level of security.

PERCEPTION

- Perception is a process to interpret, acquire, select and then organize the sensory information that is captured from the real world.
  For example: Human beings have sensory receptors such as touch, taste, smell, sight and hearing. So, the information received from these receptors is transmitted to human brain to organize the received information.
- According to the received information, action is taken by interacting with the environment to manipulate and navigate the objects.
- Perception and action are very important concepts in the field of Robotics. The following figures show the complete **autonomous robot.**

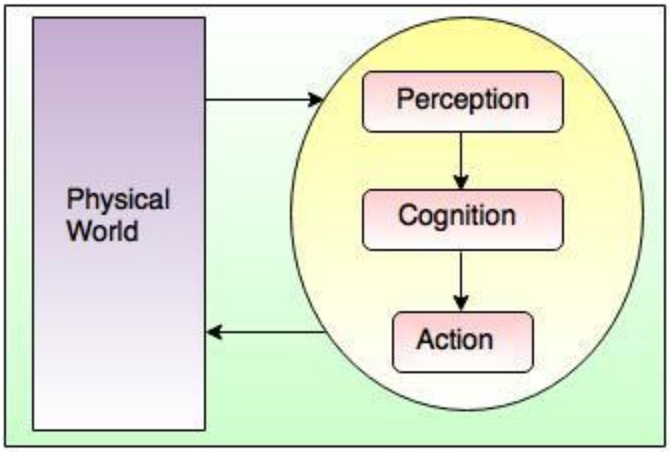


Fig: Autonomous Robot

- There is one important difference between the artificial intelligence program and robot. The AI program performs in a computer stimulated environment, while the robot performs in the physical world.

  Example:
  In chess, an AI program can be able to make a move by searching different nodes and has no facility to touch or sense the physical world.
  However, the chess playing robot can make a move and grasp the pieces by interacting with the physical world.

Image formation

Image formation is a physical process that captures object in the scene through lens and creates a 2-D image.

Image formation in digital camera

Let's understand the geometry of a pinhole camera shown in the following diagram.

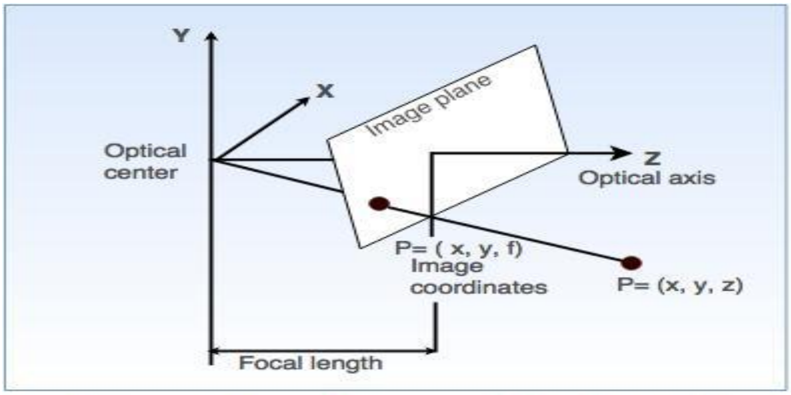


Fig: Geometry of Image Formation in the pinhole camera

In the above figure, an optical axis is perpendicular to the image plane and image plane is generally placed in front of the optical center.
So, let **P** be the point in the scene with coordinates (X,Y,Z) and **P'** be its image plane with coordinates (x, y, z).

If the focal length from the optical center is **f**, then by using properties of similar triangles, equation is derived as,

-x/f = X/Z so x = - fX/Z ....equation (i)
-y/f = -Y/Z so y = - fY/Z ....equation (ii)

These equations define an image formation process called as **perspective projection.**

What is the purpose of edge detection?

- Edge detection operation is used in an image processing.
- The main goal of edge detection is to construct the ideal outline of an image.
- Discontinuity in brightness of image is affected due to:
  i) Depth discontinuities
  ii) Surface orientation discontinuities

iii) Reflectance discontinuities
iv) Illumination.

3D-Information extraction using vision

The 3-D information extraction process plays an important role to perform the tasks like manipulation, navigation and recognition. It deals with the following aspects:

1. Segmentation of the scene

The segmentation is used to arrange the array of image pixels into regions. This helps to match semantically meaningful entities in the scene.

- The goal of segmentation is to divide an image into regions which are homogeneous.
- The union of the neighboring regions should not be homogeneous.
- **Thresholding** is the simplest technique of **segmentation.** It is simply performed on the object, which has an homogeneous intensity and a background with a different intensity level and the pixels are partitioned depending on their intensity values.
-

2. To determine the position and orientation of each object

Determination of the position and orientation of each object relative to the observer is important for manipulation and navigation tasks.
For example: Suppose a person goes to a store to buy something. While moving around he must know the locations and obstacles, so that he can make the plan and path to avoid them.

- The whole orientation of image should be specified in terms of a three dimensional rotation.
-

3. To determine the shape of each and every object

When the camera moves around an object, the distance and orientation of that object will change but it is important to preserve the shape of that object.
For example: If an object is cube, that fact does not change, but it is difficult to represent the global shape to deal with wide variety of objects present in the real world.

- If the shape of an object is same for some manipulating tasks, it becomes easy to decide how to grasp that object from a particular place.
- The **object recognition** plays most significant role to identify and classify the objects as an example only when the geometric shapes are provided with color and texture

There are number of techniques available in the visual stimulus for 3D-image extraction such as **motion, binocular stereopsis, texture, shading, and contour**. Each of these techniques operates on the background assumptions about physical scene to provide interpretation.

Image processing is the process of transforming an image into a digital form and performing certain operations to get some useful information from it. The image processing system usually treats all images as 2D signals when applying certain predetermined signal processing methods.

There are five main types of image processing:

- Visualization - Find objects that are not visible in the image
- Recognition - Distinguish or detect objects in the image
- Sharpening and restoration - Create an enhanced image from the original image
- Pattern recognition - Measure the various patterns around the objects in the image
- Retrieval - Browse and search images from a large database of digital images that are similar to the original image

Fundamental Image Processing Steps

Image Acquisition

Image acquisition is the first step in image processing. This step is also known as preprocessing in image processing. It involves retrieving the image from a source, usually a hardware-based source.

Image Enhancement

Image enhancement is the process of bringing out and highlighting certain features of interest in an image that has been obscured. This can involve changing the brightness, contrast, etc.

Image Restoration

Image restoration is the process of improving the appearance of an image. However, unlike image enhancement, image restoration is done using certain mathematical or probabilistic models.

Color Image Processing

Color image processing includes a number of color modeling techniques in a digital domain. This step has gained prominence due to the significant use of digital images over the internet.

Wavelets and Multiresolution Processing

Wavelets are used to represent images in various degrees of resolution. The images are subdivided into wavelets or smaller regions for data compression and for pyramidal representation.

Compression

Compression is a process used to reduce the storage required to save an image or the bandwidth required to transmit it. This is done particularly when the image is for use on the Internet.

Morphological Processing

Morphological processing is a set of processing operations for morphing images based on their shapes.

Segmentation

Segmentation is one of the most difficult steps of image processing. It involves partitioning an image into its constituent parts or objects.

Representation and Description

After an image is segmented into regions in the segmentation process, each region is represented and described in a form suitable for further computer processing. Representation deals with the image’s characteristics and regional properties. Description deals with extracting quantitative information that helps differentiate one class of objects from the other.

RECOGNITION

Recognition assigns a label to an object based on its description.

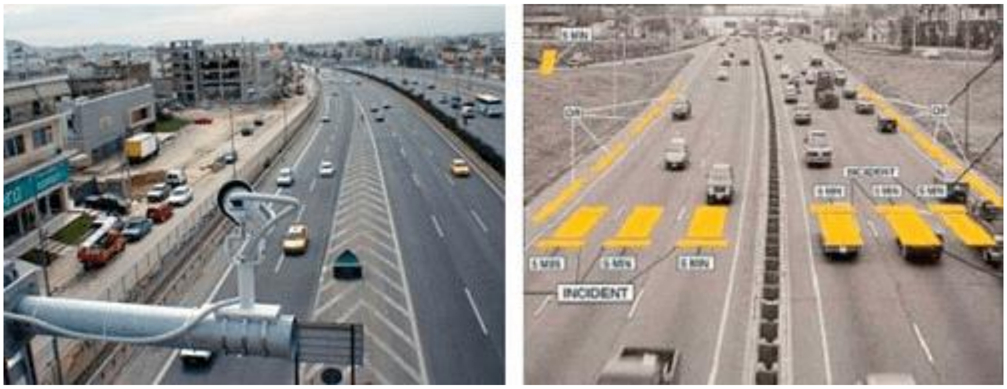
Applications of Image Processing

Medical Image Retrieval

Image processing has been extensively used in medical research and has enabled more efficient and accurate treatment plans. For example, it can be used for the early detection of breast cancer using a sophisticated nodule detection algorithm in breast scans. Since medical usage calls for highly trained image processors, these applications require significant implementation and evaluation before they can be accepted for use.

Traffic Sensing Technologies

In the case of traffic sensors, we use a video image processing system or VIPS. This consists of a) an image capturing system b) a telecommunication system and c) an image processing system. When capturing video, a VIPS has several detection zones which output an "on" signal whenever a vehicle enters the zone, and then output an "off" signal whenever the vehicle exits the detection zone. These detection zones can be set up for multiple lanes and can be used to sense the traffic in a particular station.



Left - normal traffic image | Right - a VIPS image with detection zones (source)

Besides this, it can auto record the license plate of the vehicle, distinguish the type of vehicle, monitor the speed of the driver on the highway and lots more.

Image Reconstruction

Image processing can be used to recover and fill in the missing or corrupt parts of an image. This involves using image processing systems that have been trained extensively with existing photo datasets to create newer versions of old and damaged photos.

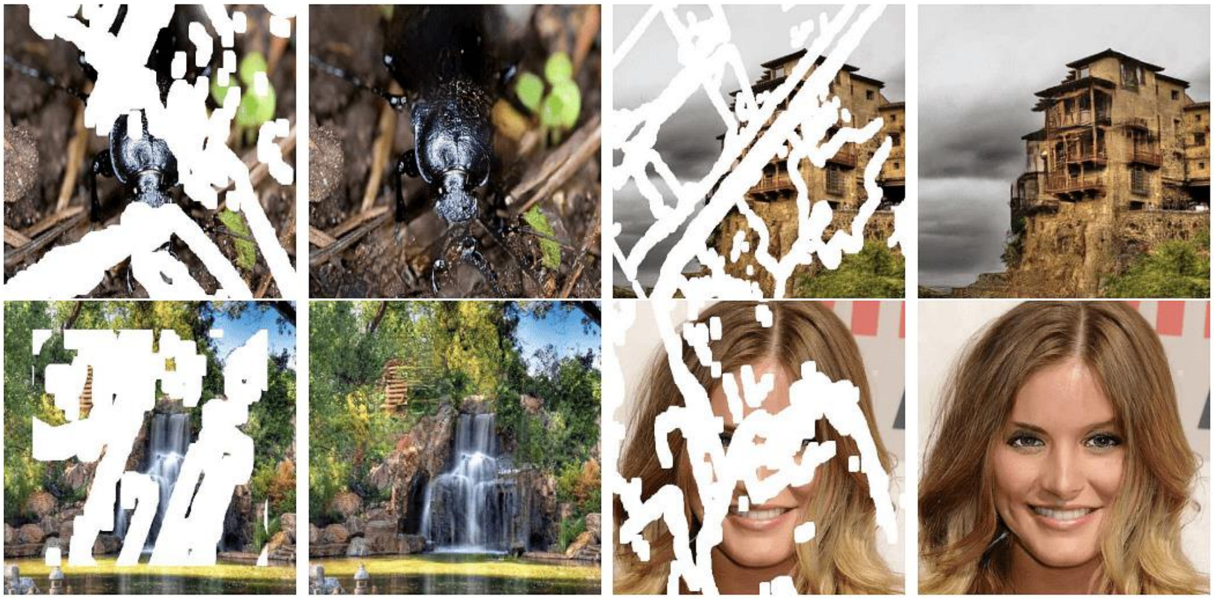


Fig: Reconstructing damaged images using image processing

Face Detection

One of the most common applications of image processing that we use today is face detection. It follows deep learning algorithms where the machine is first trained with the specific features of human faces, such as the shape of the face, the distance between the eyes, etc. After teaching the machine these human face features, it will start to accept all objects in an image that resemble a human face. Face detection is a vital tool used in security, biometrics and even filters available on most social media apps these days.

Benefits of Image Processing

The implementation of image processing techniques has had a massive impact on many tech organizations. Here are some of the most useful benefits of image processing, regardless of the field of operation:

- The digital image can be made available in any desired format (improved image, X-Ray, photo negative, etc)
- It helps to improve images for human interpretation
- Information can be processed and extracted from images for machine interpretation
- The pixels in the image can be manipulated to any desired density and contrast
- Images can be stored and retrieved easily
- It allows for easy electronic transmission of images to third-party providers

There are two methods of image processing:

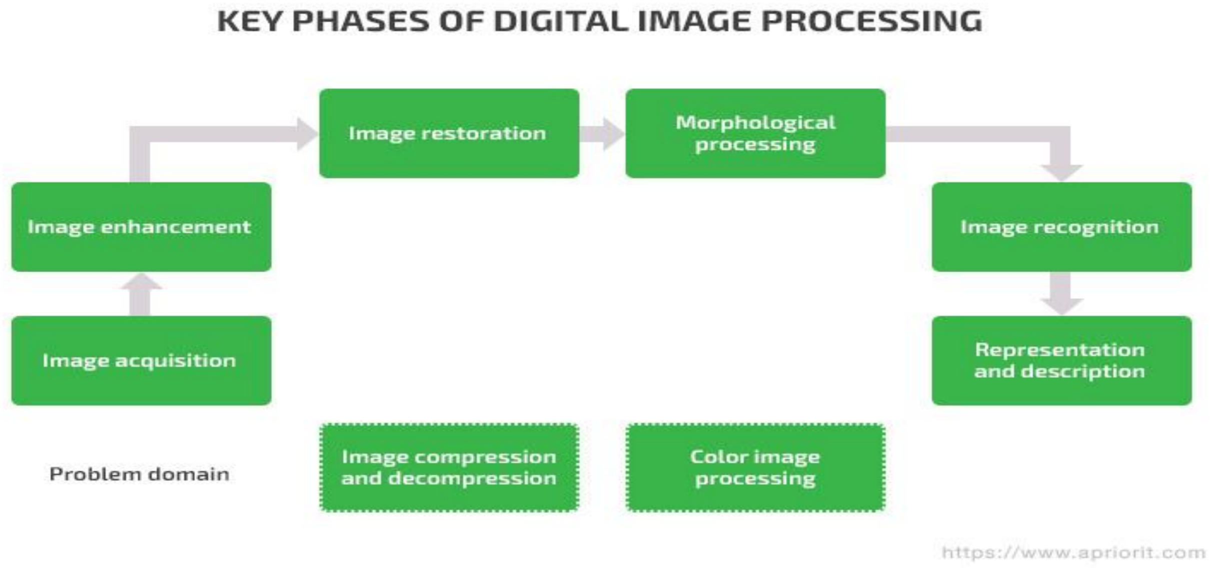
Analog Image processing:It is used for processing physical photographs,printouts and other hard copies of images

For analog image processing, the output is always an image.

Digital Image Processing: It is used for manipulating digital images with the help of computer algorithms

For digital image processing, however, the output may be an image or information associated with that image, such as data on features, characteristics, bounding boxes, or masks

Digital image processing includes eight key phases:



1. **Image acquisition** is the process of capturing an image with a sensor (such as a camera) and converting it into a manageable entity (for example, a digital image file). One popular image acquisition method is scraping.

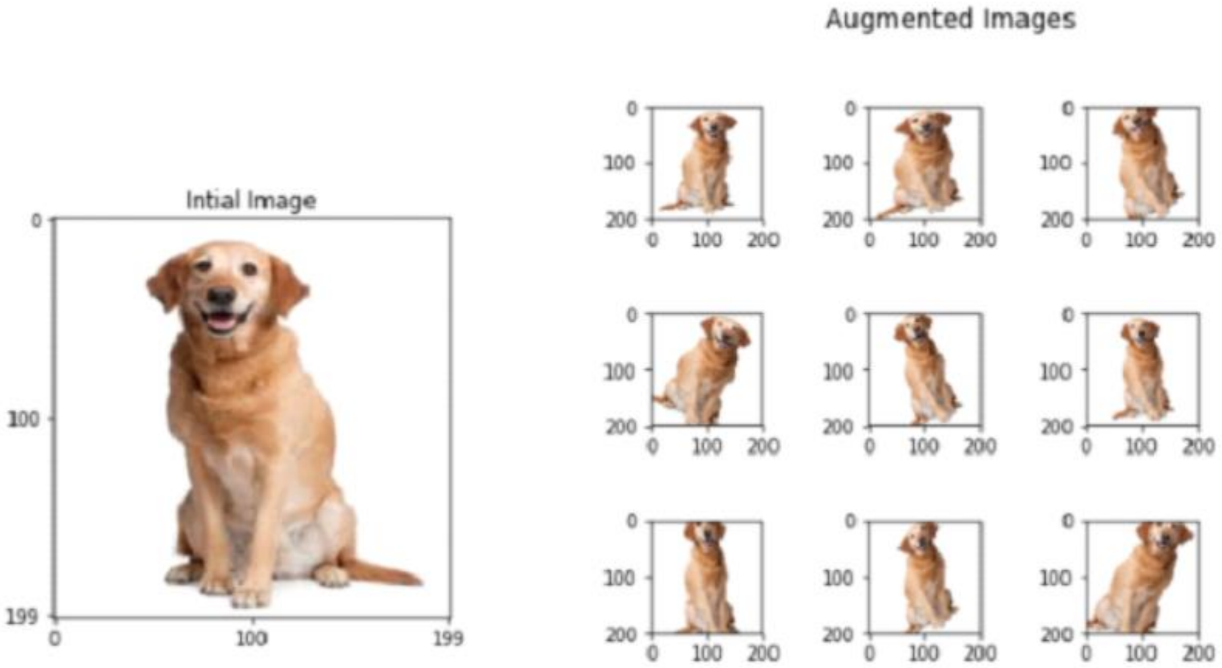
2. **Image enhancement** improves the quality of an image in order to extract hidden information from it for further processing.

3. **Image restoration** also improves the quality of an image, mostly by removing possible corruptions in order to get a cleaner version. This process is based mostly on probabilistic and mathematical models and can be used to get rid of blur, noise, missing pixels, camera misfocus, watermarks, and other corruptions that may negatively affect the training of a neural network.

4. **Color image processing** includes the processing of colored images and different color spaces. Depending on the image type, we can talk about pseudocolor processing (when colors are assigned grayscale values) or RGB processing (for images acquired with a full-color sensor).

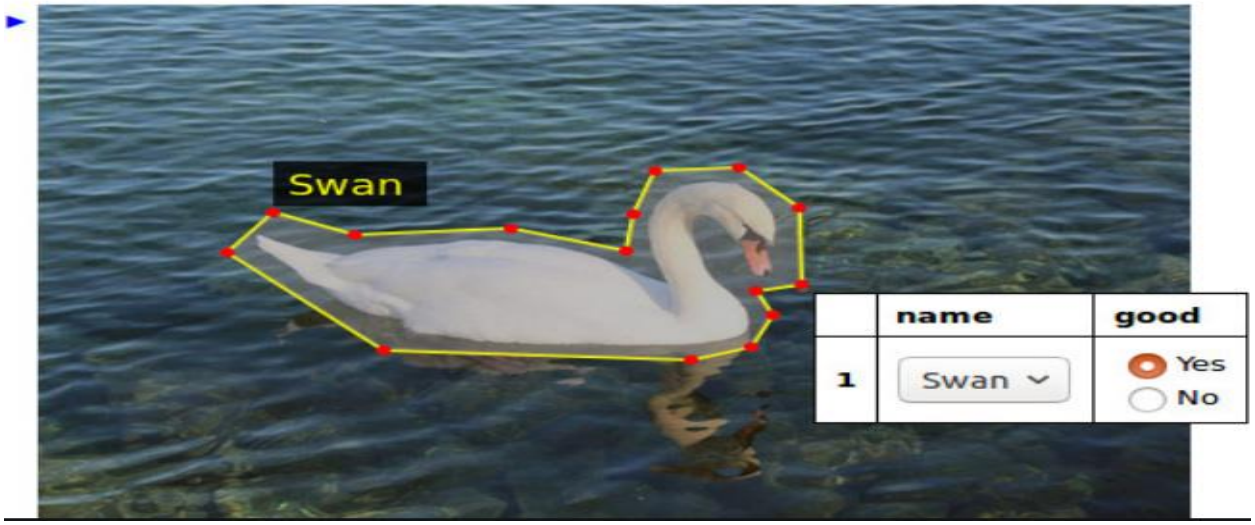
5. **Image compression and decompression** allow for changing the size and resolution of an image. Compression is responsible for reducing the size and resolution, while decompression is used for restoring an image to its original size and resolution.

These techniques are often used during the image augmentation process. When you lack data, you can extend your dataset with slightly augmented images. In this way, you can improve the way your neural network model generalizes data and make sure it provides high-quality results.



Example: Image Augumentation

6. **Morphological processing** describes the shapes and structures of the objects in an image. Morphological processing techniques can be used when creating datasets for training AI models. In particular, morphological analysis and processing can be applied at the annotation stage, when you describe what you want your AI model to detect or recognize.



	name	good
1	Swan ⌄	● Yes ○ No

An example of the annotation process of morphological analysis

7. **Image recognition** is the process of identifying specific features of particular objects in an image. AI-based image recognition often uses such techniques as object detection, object recognition, and segmentation.

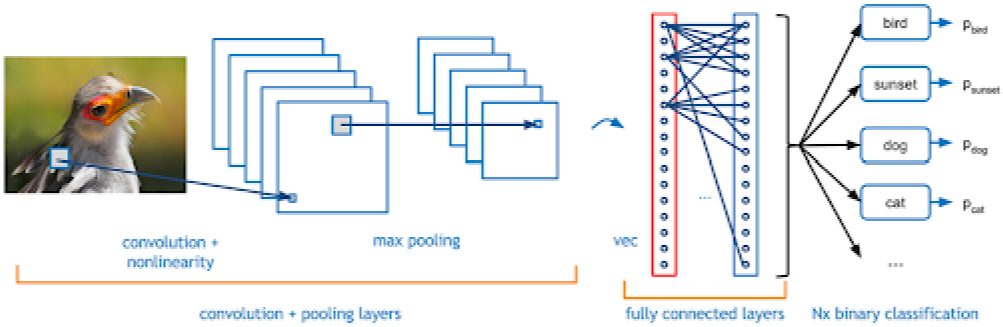


Image recognition with a CNN

8. **Representation and description** is the process of visualizing and describing processed data. AI systems are designed to work as efficiently as possible. The raw output of an AI system looks like an array of numbers and values that represent the information the AI model was trained to produce. Yet for the sake of system performance, a deep neural network usually doesn’t include any output data representations. Using special visualization tools, you can turn these arrays of numbers into readable images suitable for further analysis.

OBJECT RECOGNITION

Object recognition is a computer vision technique for identifying objects in images or videos. Object recognition is a key output of deep learning and machine learning algorithms. When humans look at a photograph or watch a video, we can readily spot people, objects, scenes, and visual details. The goal is to teach a computer to do what comes naturally to humans: to gain a level of understanding of what an image contains.

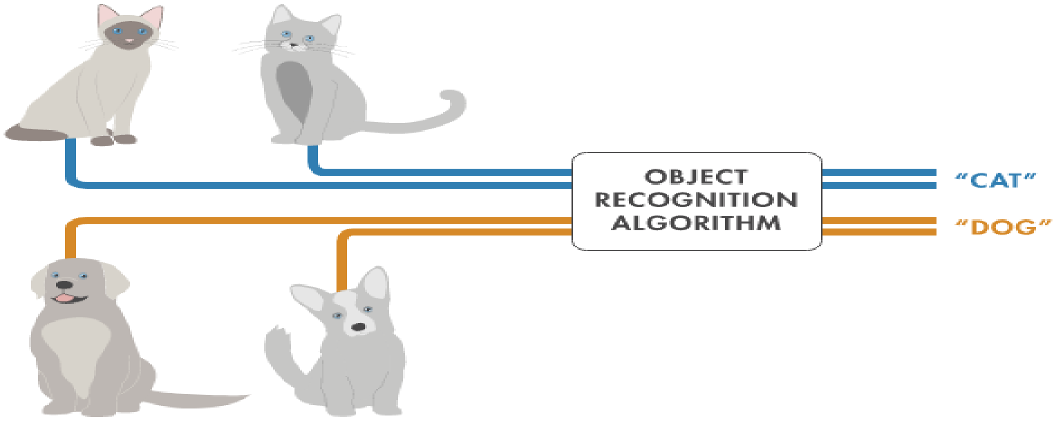


Figure 1. Using object recognition to identify different categories of objects.

Object recognition is a key technology behind driverless cars, enabling them to recognize a stop sign or to distinguish a pedestrian from a lamppost. It is also useful in a variety of applications such as disease identification in bioimaging, industrial inspection, and robotic vision.

Object Recognition vs. Object Detection

Object detection and object recognition are similar techniques for identifying objects, but they vary in their execution. Object detection is the process of finding instances of objects in images. In the case of deep learning, object detection is a subset of object recognition, where the object is not only identified but also located in an image. This allows for multiple objects to be identified and located within the same image.

How Object Recognition Works

You can use a variety of approaches for object recognition. Recently, techniques in machine learning and deep learning have become popular approaches to object recognition problems. Both techniques learn to identify objects in images, but they differ in their execution.

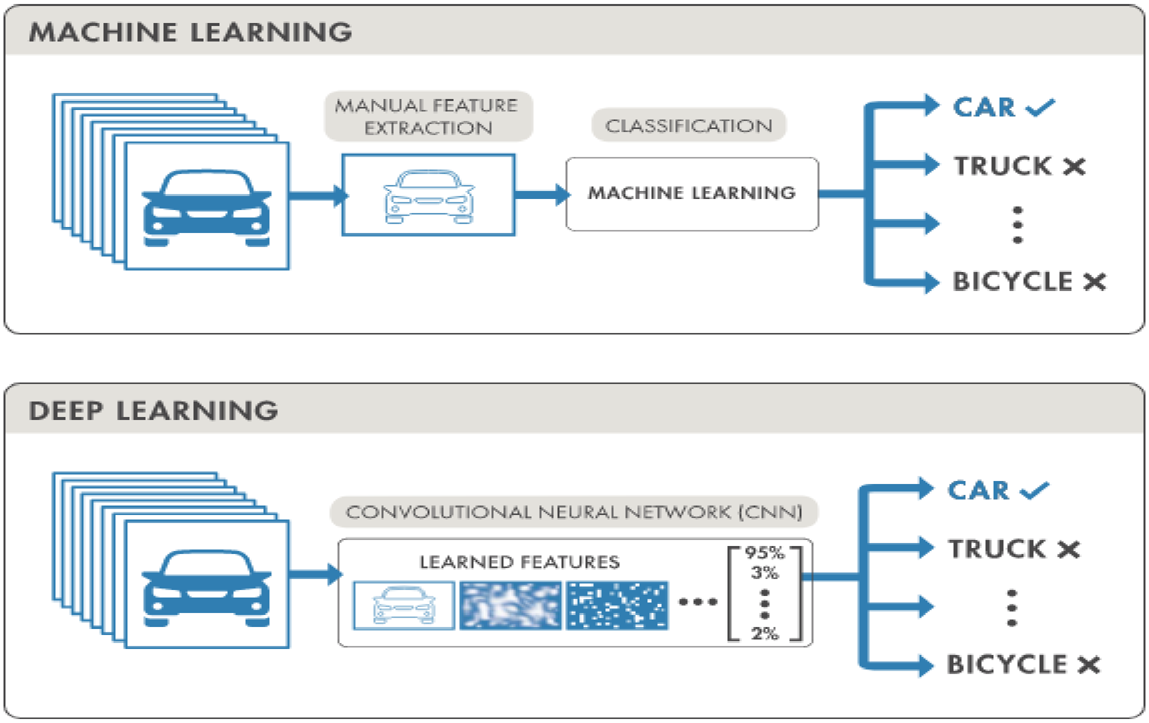


Figure 3: Machine learning and deep learning techniques for object recognition.

OBJECT RECOGNITION TECHNIQUES

1.Template matching

Template matching is a technique for finding small parts of an image which match a template image. It is a straightforward process. In this technique template images for different objects are stored. When an image is given as input to the system, it is matched with the stored template images to determine the object in the input image. Templates are frequently used for recognition of characters, numbers, objects, etc. It can be performed on either color or gray level images. Template matching can either be pixel to pixel matching or feature based. In feature based the features of template image is compared to features of sub-images of the given input image; to determine if the template object is present in the input image.

B. Color based

Color provides potent information for object recognition. A simple and efficient object detection scheme is to represent and match images on the basis of color histograms.

The color information is extended in two existing methods for object detection, the part-based detection framework and the Efficient Subwindow Search approach . The three main criteria which should be taken into account when choosing an approach to integrating color into object detection are feature Combination, photometric invariance and compactness.

variety of color models used for recognition of multicolored objects according to the following criteria: 1. Robustness to a change in viewing direction

2. Robustness to a change in object geometry

3. Robustness to a change in the direction of the illumination

4. Robustness to a change in the intensity of the illumination

5. Robustness to a change in the spectral power distribution (SPD) of the illumination.

The color models have High discriminative power; robustness to object occlusion and cluttering; robustness to noise in the images.

3.Active and Passive

Object detection in passive manner does not involve local image samples extracted during scanning. Two main object-detection approaches that employ passive scanning:

1. The window-sliding approach:

It uses passive scanning to check if the object is present or not at all locations of an evenly spaced grid. This approach extracts a local sample at each grid point and classifies it either as an object or as a part of the background

2.The part-based approach:

It uses passive scanning to determine interest points in an image. This approach calculates an interest value for local samples at all points of an evenly spaced grid. At the interest points, the approach extracts new local samples that are evaluated as belonging to the object or the background

Some methods try to bound the region of the image in which passive scanning is applied. It is a computationally expensive and inefficient scanning method. In this method at each sampling point costly feature extraction is performed, while the probability of detecting an object or suitable interest point can be squat In active scanning local samples are used to guide the scanning process. At the current scanning position a local image sample is extracted and mapped to a shifting vector indicating the next scanning position. The method takes successive samples towards the expected object location, while skipping regions unlikely to contain the object. The goal of active scanning is to save computational effort, while retaining a good detection performance.

The active object-detection method (AOD-method) scans the image for multiple discrete time steps in order to find an object. In the AOD-method this process consists of three phases:

1.Scanning for likely object locations on a coarse scale

2. Refining the scanning position on a fine scale

3.Verifying object presence at the last scanning position with a standard object detector.

4.Shape based

Recently, shape features have been extensively explored to detect objects in real-world images. The shape features are more striking as compared to local features like SIFT because most object categories are better described by their shape then texture, such as cows, horses and cups and also for wiry objects like bikes, chair or ladders, local features unavoidably contain large amount of background mess. Thus shape features are often used as a replacement or complement to local features

Berg, et.al. have proposed a new algorithm to find correspondences between feature points for object recognition in the framework of deformable shape matching. The basic subroutine in deformable shape matching takes as input an image with an unknown object (shape) and compares it to a model by solving the correspondence problem between the model and the object. Then it performs aligning transformation and computes a similarity based on both the aligning transform and the residual after applying the aligning transformation

5. Local and global features

The most common approach to generic object detection is to slide a window across the image and to classify each such local window as containing the target or background. This approach has been successfully used to detect rigid objects such as faces and cars in and In, a method of object recognition and segmentation using Scale-Invariant Feature Transform (SIFT) and Graph Cuts is presented. SIFT feature is invariant for rotations, scale changes, and illumination changes. By combing SIFT and Graph Cuts, the existence of objects is recognized first by vote processing of SIFT keypoints. Then the object region is cut out by Graph Cuts using SIFT keypoints as seeds. Both recognition and segmentation are performed automatically under cluttered backgrounds including occlusion.

Object Recognition Using Deep Learning

Deep learning techniques have become a popular method for doing object recognition. Deep learning models such as convolutional neural networks, or CNNs, are used to automatically learn an object's inherent features in order to identify that object. For example, a CNN can learn to identify differences between cats and dogs by analyzing thousands of training images and learning the features that make cats and dogs different. There are two approaches to performing object recognition using deep learning:

- **Training a model from scratch**: To train a deep network from scratch, you gather a very large labeled dataset and design a network architecture that will learn the features and build the model. The results can be impressive, but this approach requires a large amount of training data, and you need to set up the layers and weights in the CNN.
- **Using a pretrained deep learning model**: Most deep learning applications use the transfer learning approach, a process that involves fine-tuning a pretrained model. You start with an existing network, such as AlexNet or GoogLeNet, and feed in new data containing previously unknown classes. This method is less time-consuming and can provide a faster outcome because the model has already been trained on thousands or millions of images.

Deep learning offers a high level of accuracy but requires a large amount of data to make accurate predictions.

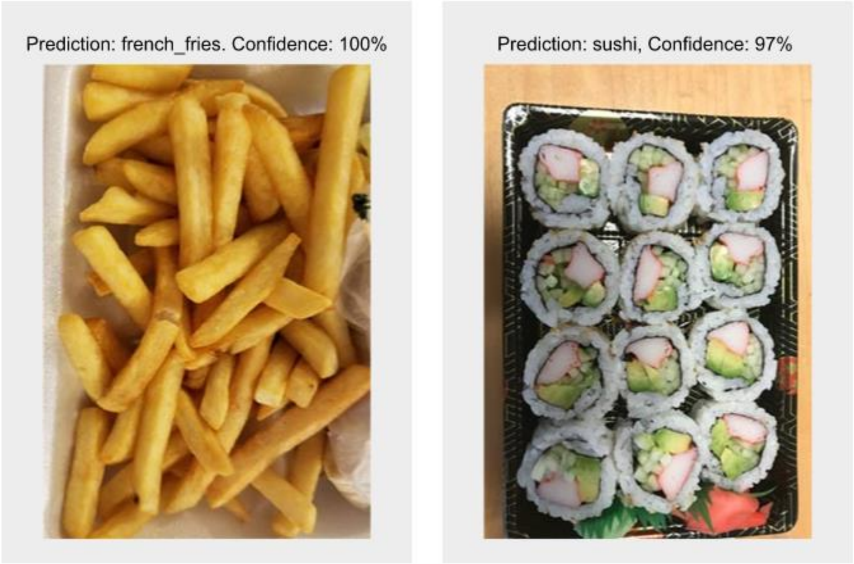

Prediction: french_fries. Confidence: 100%
Prediction: sushi, Confidence: 97%

Figure 4: Deep learning application showing object recognition of restaurant food.

Machine Learning Workflow

To perform object recognition using a standard machine learning approach, you start with a collection of images (or video), and select the relevant features in each image. For example, a feature extraction algorithm might extract edge or corner features that can be used to differentiate between classes in your data.

These features are added to a machine learning model, which will separate these features into their distinct categories, and then use this information when analyzing and classifying new objects.

You can use a variety of machine learning algorithms and feature extraction methods, which offer many combinations to create an accurate object recognition model.

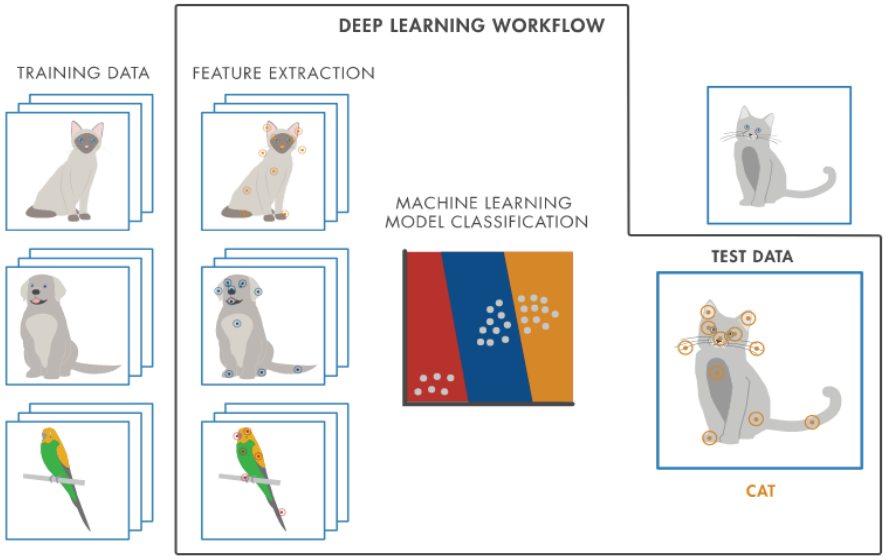


Figure 5: Machine learning workflow for object recognition.

OBJECT RECOGNITION BY APPEARANCE

Appearance means what an object tends to look like. For example, football is rather round in shape. It is important to know every class of images with a classifier. Taking the example of faces, looking at the camera-every face looks similar under good light and perfect resolution. A strategy called sliding window includes computing features for an object and present it to a classifier. One strategy is to estimate and correct the illumination in each image window and another is to build features out of gradient orientations. To find faces of different sizes, repeat the sweep over larger or smaller versions of the image. Then, we post process the responses across scales and location to produce the final set of detections. Postprocessing have several overlapping windows that each report a match for a face. To yield a single high quality match, we can combine these partial overlapping matches at nearby locations. Therefore, it gives a face detector that can search over locations and scales.

1.Complex appearance and pattern elements

Since, several effects can move features around in an image of the object, many objects produce much more complex patterns than faces do. Effects include: Foreshortening: which causes a pattern viewed at a slant to be significantly distorted Aspect: which causes objects to look different when seen from different directions. Occlusion: when some parts are hidden from some viewing directions. Self-occlusion can be defined as objects can occlude one another or parts of an object can occlude other parts. Deformation: where internal degrees of freedom of the object change its appearance. An object recognizer is then a collection of features that can tell whether the pattern elements are present, and whether they are in about the right place. With a histogram of the pattern elements that appear can be considered as the most obvious approach to represent the image window. This approach does not work particularly well, because too many patterns get confused with one another.

2. Pedestrian detection with HOG features

Each year car accidents kill about 1.2 million people, and to avoid this problem cars should be provided with sensors which detect pedestrians and result in saving many lives. The most usual cases are lateral or frontal views of a walk. In these cases, we see either a “lollipop” shape-the torser is wider than the legs, which are together in the stance phase of the walk-or a “scissor” shape-where the legs are swinging in the walk. Therefore, we need to build a useful movingwindow pedestrian detector. To represent the image window, it is better to use orientations than edges because there isn’t always a strong contrast between a pedestrian and the background. Pedestrians can move their arms and legs around, so we should use a histogram to suppress some spatial detail in the feature. When breaking up the window into cells, overlapping occurs and hence, build an orientation histogram in each cell. Through this feature, we can determine whether head-and-shoulders curve is at the top of the window or at the bottom, but will not change, if the head is moved bit slightly.

RECONSTRUCTING THE 3D WORLD

This section shows how to go from 2D image to a 3D representation of the scene. A fundamental question arises which is how do we recover 3D information when given all the points in the scene that fall along a ray to pinhole are projected to the same point in the image? The two solutions are, If we have two or more images from different• camera positions, then we can triangulate to find the position of a point in the scene. We can exploit background knowledge about the• physical scene that gave rise to the image. Given an object model P(scene) and a redering model P(image |scene).We can compute a posterior distribution P(scene|image). For a scene reconstruction, we survey 8 commonly used visual cues because there is no single unified theory for it. Motion Parallax: We state an equation understanding the concept of relation among the optical flow velocity and depth in the scene.

$$v_x(x,y) = \frac{-T_x + xT_z}{Z(x,y)}, \qquad v_y(x,y) = \frac{-T_y + yT_z}{Z(x,y)},$$

The point in the scene corresponding to the point in the image at (x,y).Both components of the optical flow,vx(x,y) and vy(x,y), are zero at the point x=Tx/Tz ,y=Ty/Tz .This point is called focus of expansion of the flow field. If we change the origin in the x-y plane to lie at the focus of expansion, then the expressions for optical flow take on a particularly simple form. Let (x’,y’) be the new coordinates defined by x’=x-Tx/Tz ,y’=y-Ty/Tz. Then,

$$v_x(x',y') = \frac{x'T_z}{Z(x',y')}, \qquad v_y(x',y') = \frac{y'T_z}{Z(x',y')} .$$

Binocular Stereopsis

This idea is similar to motion parallax, but we use two or more images separated in space. Binocular stereopsis enables a wilder field of vision for the predators who have eyes in the front. If we super pose the two images there will be a disparity in the location of the image feature in the two images as a given feature in the scene will be in a different place relative to the Z-axis of each image plane. Using optical flow equations vector T acting for time δt, with Tx=b/ δt and Ty=Tz=0. Horizontal disparity is equal to the ratio of base line to the depth, and vertical disparity is zero i.e,H=b/Z,V=0. Humans fixate at a point in the scene at which the optical axes of the two eyes intersect nder normal viewing conditions. The actual disparity is δθ, we have disparity=bδZ/Z2. In humans, b(the base line distance between the eyes) is about 6cm. So for Z=30cm we get small value δZ=0.036mm which means at a distance of 30cm humans can differentiate the depths which differ by a little length as 0.036mm.

Multiple views

Most of the techniques that have been developed had make the use of the information available in multiple views, even from hundreds or thousands of cameras. There are few problems in multiple views which can be solved algorithmically: Correspondence problem: In the 3D world,• identifying features in the different images that are projections of the same feature. Relative orientation problem: Finding out the• transformation between the coordinate systems fixed to the different cameras. Depth estimation problem: Finding out the depths• of various points in the world for which image plane projections were available in at least two views.

Texture

Texture is used to estimate distances and for segmenting objects. The texture elements are also known as texels.

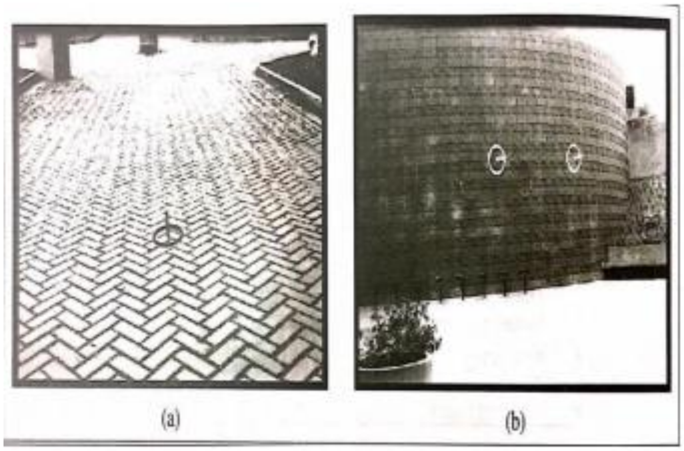


Fig.2. Texture

As shown in the above figure, the paving tiles are identical in the scene but appear different in the image for 2 reasons. Differences in the distances of the texels from the• camera. Differences in the foreshortening of the texels.• Various algorithms have been developed to make use of variation in the appearance of the projected texels as a basis for finding out the surface normals but they were not accurate as much as the algorithms which were used for multiple views.

Mathematical description of reconstruction

Given a group of 3D points viewed by N cameras with matrices $\{P^i\}_{i=1\ldots N}$, define $m_j^i \simeq P^i w_j$. , to be the homogeneous coordinates of the projection of the **Jth** point onto the **ith** camera. The reconstruction problem can be changed to: given the group of pixel coordinates $\{m_j^i\}$, find the corresponding set of camera matrices $\{P^i\}$ and the scene structure

$\{w_j\}$ such that

$$m_j^i \simeq P^i w_j \quad (1)$$

Generally, without further restrictions, we will obtain a projective reconstruction.[4][5] If $\{P^i\}$ and $\{w_j\}$ satisfy (1), $\{P^iT\}$ and $\{T^{-1}w_j\}$ and will satisfy (1) with any **4 × 4** nonsingular matrix **T**.

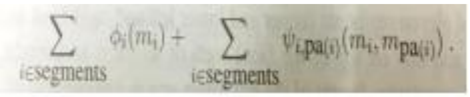
A projective reconstruction can be calculated by correspondence of points only without any a priori information

OBJECT RECOGNIZATION FROM STRUCTURAL INFORMATION

Human body parts are very small in images and vary in color, texture among individuals and hence it is difficult to detect them using moving window method. Some parts are very small to a size of two to three pixels wide. As the layout of body describes the movement, it is important to conclude layout of body in the images. whether the configuration are acceptable or not it can be described by a model called deformable template. The simplest deformable template model of a person connects lower arms to upper arms, upper arms to torso, and so on.

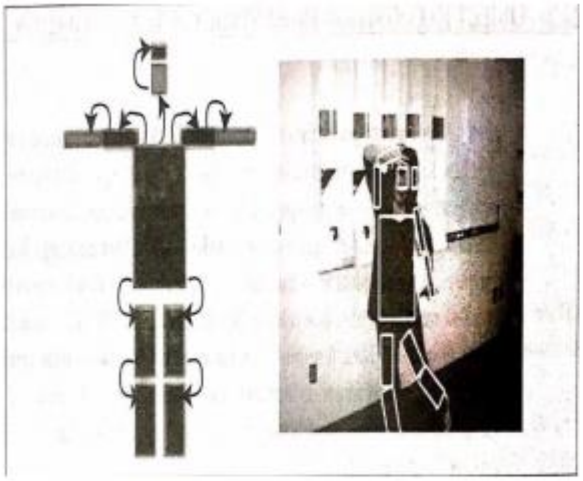
1. The geometry of bodies: finding arms and legs

A tree of eleven segments is used to model the geometry of body. Segments are rectangular in shape.” Cardboard people” are few models which are used to assume the position and pose(orientation) of human body parts and segments in the images. Evaluation of configuration is done based on 2 criteria: First, an image rectangle should look like its segment. A function φi describes how well an image rectangle matches a body segment. Function ψ defines how the relations of a pair of rectangle segments meet the expected body segments. Each segment has only one parent because the dependencies between segments form a tree. The parent can be described by a function ψi,p,a(i).



$$\sum_{i\in\text{segments}} \phi_i(m_i) + \sum_{i\in\text{segments}} \psi_{i,\mathrm{pa}(i)}(m_i, m_{\mathrm{pa}(i)}) \,.$$

There is an angle between segments mainly for the ankles and knees to be differentiated. If there are M image rectangles having the right torso as O(M6) for a model, then the best allocation of rectangles to segments will get slow. However this can be solved by using various speed-ups which are available for an appropriate choice of ψ. This model is usually known as pictorial structure model.



We generally have to build a model of segment appearances when we don't have an idea of what a person looks like. Appearance model is the description of what a person looks like.

2 .Coherent appearance: tracking people in video

By improving the concept of tracking people in a video we can get further in game interfaces and surveillance systems, but many methods haven't succeeded with the problem because people in the videos tend to move fast and produce large accelerations. The effective methods state the fact that, from frame to frame the appearances change. We assume the video to be a huge collection of pictures we wish to track. The people in the video can be tracked by detecting their body segments. In some cases, the segment detector may generate lots of false positives because the people are appearing against a near fixed background. This problem can be solved using an alternative which is quite in practice, by applying a detector to a fixed body configuration for all the frames. We can know when we found a real person in a video by tuning the detector to low false positive rate.

<u>USING VISION</u>

If vision systems could analyze video and understood what people are doing, we would be able to: design buildings and public places better by collecting and using data about what people do in public; build more accurate, more secure, and less intrusive surveillance systems; build computer sports commentators; and build human-computer interfaces that watch people and react to their behavior

Using vision for controlling movement

One of the principal uses of vision is to provide information both for manipulating objects— picking them up, grasping them, twirling them, and so on—and for navigating while avoiding obstacles. The ability to use vision for these purposes is present in the most primitive of animal visual systems. In many cases, the visual system is minimal, in the sense that it extracts from the available light field just the information the animal needs to inform its behavior. Quite probably, modern vision systems evolved from early, primitive organisms that used a photosensitive spot at one end to orient themselves toward (or away from) the light.

Let us consider a vision system for an automated vehicle driving on a freeway. The tasks faced by the driver include the following:

1. **Lateral control**—ensure that the vehicle remains securely within its lane or changes lanes smoothly when required.

2. **Longitudinal control**—ensure that there is a safe distance to the vehicle in front.

3. **Obstacle avoidance**—monitor vehicles in neighboring lanes and be prepared for evasive maneuvers if one of them decides to change lanes. The problem for the driver is to generate appropriate steering, acceleration, and braking actions to best accomplish these tasks.

For lateral control, one needs to maintain a representation of the position and orientation of the car relative to the lane. We can use edge-detection algorithms to find edges corresponding to the lane-marker segments. We can then fit smooth curves to these edge elements. The parameters of these curves carry information about the lateral position of the car, the direction it is pointing relative to the lane, and the curvature of the lane. This information, along with information about the dynamics of the car, is all that is needed by the steering-control system. If we have good detailed maps of the road, then the vision system serves to confirm our position (and to watch for obstacles that are not on the map).

For longitudinal control, one needs to know distances to the vehicles in front. This can be accomplished with binocular stereopsis or optical flow. Using these techniques, visioncontrolled cars can now drive reliably at highway speeds.