Computer Vision for Image Understanding: A Comprehensive Review

[1]Diksha Jha
[1,]*Department of Mathematics,Birla Institute of Technology, Mesra, Ranchi, India*

A survey on computer vision for image understanding

Abstract. In computer vision we have our own Turing test. Can a machine describe the contents of an image or a video in the way a human being would do? We want to know how far we have advanced in image recognition. In recent years, Deep Learning has increased considerably the precision rate of many tasks related to computer vision. Many datasets of labeled images are now available online, which leads to pre-trained models for many computer vision applications. In this work, we gather information of the latest techniques to perform image understanding and description. As a conclusion we obtained that the combination of Natural Language Processing (using Recurrent Neural Networks and Long Short-Term Memory) plus Image Understanding (using Convolutional Neural Networks) could bring new types of powerful and useful applications in which the computer will be able to answer questions about the content of images and videos. In order to build datasets of labeled images, it is needed a lot of work and most of the datasets are constructed using crowd work. These new applications have the potential to increase the human machine interaction to new levels of usability and user' satisfaction

Keywords: Computer Vision, Deep Learning, Image Understanding, CNN, Scene Recognition, Object classification.

Introduction

Computer vision is a subfield of Artificial Intelligence in which they try to mimic the function of the visual system of human beings; that indeed is a very complex task. With the exponential growth of the computing power and the increased number of cameras that are installed all over the cities, now it is possible to build automated systems that accomplish with computer vision tasks. However, complex tasks (HIT = Human Intelligence Task) are still in investigation. Among these difficult tasks, we have image understanding in which the computer is able to describe the image in a similar way a human being would describe the image. Due to its complexity and its potential, it is worth researching this area for new applications. Some examples of Image Understanding applications are describe in Table 1:

Another interesting research direction could be to transfer knowledge of common objects to learn rare object models.

Table 1. Image Understanding Applications.

Area	Example
Automation of Industrial Processes	- Object acquisition by robot arms, for example by "bin picking." - Automatic guidance of seam welders and cutting tools. - Very large scale integrated circuit-(VSLI-) related processes, such as

	lead bonding, chip alignment, and packaging. - Monitoring, filtering, and thereby containing the flood of data from oil drill sites or from seismographs. - Providing visual feedback for automatic assembly and repair.
Inspection tasks	- The inspection of printed circuit boards for spurs, shorts, and bad connections. - Checking the results of casting processes for impurities and fractures. - Screening of medical images such as chromosome slides, cancer smears, x-ray and ultrasound images, and tomography. - Routine screening of plant samples.
Remote Sensing	- Cartography, the automatic generation of hill shaded maps, and the registration of satellite images with terrain maps. - Monitoring of traffic along roads, docks, and at airfields. - Management of land resources such as water, forestry, soil erosion and crop growth.
Making Computer Power More Accessible	- Management information systems that have a communication channel considerably wider than current systems that are addressed by typing or pointing. - Document readers (for those that still use paper). - Design aids for architects and mechanical engineers.
Military Applications	- Tracking moving objects. - Automatic navigation based on passive sensing. - Target acquisition and range finding.
Aids for the Partially Sighted	- Systems that read a document and say what was read. - Automatic "guide dog" navigation systems.

Theoretical framework

Image Segmentation. It consists in assigning a class to every pixel of an image. In image segmentation the algorithm puts the same label to the instances of the same class, for example cars are seen with the same color.

Event recognition. An event can be said as a semantically meaningful human activity, taking place within a selected environment and containing a number of necessary objects. It can also be defined as a descriptive interpretation of the visual world for the blind. In the other hand, for best understanding images we can use the 5Ws questions: Who, Where, What, When and hoW. With event recognition we can answer 3

of the 5 questions. [1] what? - The event label, where? - The scene environment label, who? - A list of the object categories

It can be said that event recognition is composed of scene recognition + classification. The SUN dataset[2] (Scene UNderstanding) is an example of event recognition effort in which all the pictures are organized in hierarchical categories.

Scene recognition. In scene recognition, algorithms learn global statistics of the scene categories. In order to get better results and depending of the application, it is necessary to distinguish indoor scenes from outdoor scenes. In the other hand, it is known that outdoor scenes recognition models perform poorly in indoor scenes [3].

Fine-grained recognition. Is the task of distinguishing between visually very similar objects such as identifying the species of a bird, the breed of a dog or the model of an aircraft[4].

Action Classification. Its goal is to assign a label of action or event to a video. Some human actions can be recognized by a single image, but others require the motion of a person for doing that, algorithms use optical flow or dense trajectories techniques.

Examples of simple human actions are walking, running, waving, clapping.

Action Localization. It is the search of a spatial region and time interval of a specific action in a video. Action localization combines object detection, object tracking and action classification methods. It is necessary to take into account the temporal dimension and the object trajectory.

Object category recognition. Classifying images can be defined as a collection of regions, describing only their appearance and ignoring their spatial structure. For object categorization we have generative models and discriminative models. Image similarity metrics are also used for object recognition, for example, distance metrics: Dssd, Dwarp, Dshift[5]. In the other hand, object classification could be binary classification or multiclass classification.

The PASCAL object recognition challenge is an example of a competition that the goal is to get the model that best identifies objects in an image. Some of the models for object recognition are bag of words model, parts and structure models, discriminative methods, combined recognition and segmentation

BoW (Bag of Words). Bag of visual words is a vector of occurrence counts of a vocabulary of local image features. A codebook represents an image as sequence of appearance words. BoW can be treated as a supervised or unsupervised task[6]; the scene classification is a supervised task and the object discovery is unsupervised.

Hierarchical model. Objects are identified by this hierarchy:

Scene / objects / parts

Model: Parts and structure. Objects are identified by the structure:Pixels / pixels grouping / parts / objects

ConvNets. Convolutional Networks have had very good results for analyzing visual imagery, due to the fact that it tries to identify parts of objects in its convolution stages. [7]. For using ConvNets, very large amount of training data is needed. Table 2 enumerates examples of specialized Image Datasets that can be used for training a Deep Learning model.

Table 2. Specialized image databases for machine learning training.

Database	Description	Task
VOC07c	Pascal VOC 2007	Object Image Classification
VOC12c	Pascal VOC 2012	Object Image Classification
VOC12a	Pascal VOC 2012	Action Image Classification
MIT67	MIT 67 Indoor Scenes	Scene Image Classification
VOC07d	PASCAL VOC 2007	Object Detection
VOC10d	PASCAL VOC 2010	Object Detection
VOC12d	PASCAL VOC 2012	Object Detection
VOC11s	PASCAL VOC 2011	Object Category Segmentation
200Birds	UCSD-Caltech 2011-200 Birds dataset	Fine-grained Recognition
102Flowers	Oxford 102 Flowers	Fine-grained Recognition
H3Datt	H3D poselets Human 9 Attributes	Attribute Detection
UIUCatt	UIUC object attributes	Attribute Detection
LFW	Labeled Faces in the Wild	Metric Learning
Oxford5k	Oxford 5k Buildings Dataset	Instance Retrieval
Paris6k	Paris 6k Buildings Dataset[8]	Instance Retrieval
Sculp6k	Oxford Sculptures Dataset	Instance Retrieval
Holidays	INRIA Holidays Scenes Dataset	Instance Retrieval
UKB	Uni. of Kentucky Retrieval Benchmark Dataset	Instance Retrieval

Another interesting image datasets are [9] video surveillance, human health monitoring, human pose, etc., [10] 22.210 fully annotated images with objects and many with parts, [11] a database of 360 degrees panoramas, [12] a database of 400,000 spoken captions for natural images drawn from the Places 205 image dataset.

Materials and Methods

In order to get the latest scientific information, we looked for academic databases like IEEE Xplore, Google Scholar, etc. In the other hand, we found papers from 10 or 15 years ago that describe best the basic theory behind computer vision techniques and papers from recent years which describe the architecture of Convolutional Neural Networks commonly used for image understanding.

We found many models and datasets for image description which are available for the public in repositories like github; we downloaded some of them and tried to verify the benefits of those models. For experimenting with image classification, we use R Studio with the Keras package and the imagenet dataset[13]. We also use Matlab with the Deep Learning Toolbox and the AlexNet, VGG16 and VGG19 datasets.