IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Object Detection and Identification for Blinds

Sayli Jadhav (BE Student)
Department of Information Technology
Bharati Vidyapeeth College of Engineering
Navi Mumbai, India

Mandakini Bhabal (BE Student)
Department of Information Technology
Bharati Vidyapeeth College of Engineering
Navi Mumbai, India

Namrata Chavhan (BE Student)
Department of Information technology
Bharati Vidyapeeth College of Engineering
Navi Mumbai, India

Prof. V. N. Patil
Department of Information Technology
Bharati vidyapeeth college of Engineering
Navi Mumbai, India

***Abstract*—Vision plays an important role in our routine. But not for those who are visually impaired by birth, disease or due to accident. Their daily life highly disrupted due to no vision. There are so many blind people in the world who face difficulties in day to day life. They need helping hands while go out as well as identifying different objects from surrounding. There are so many devices available for visually impaired to provide navigations to them such as Electronic walking stick. But such devices only provide navigations to them but don't identify objects from surrounding. The proposed system include wearable device as main part of product. This device includes raspberry-pi model, camera module and Bluetooth earphone. The device captured real time images from surrounding using camera module. The model is trained with Mobile Net SSD (Single Shot Detector) algorithm with Pascal VOC and COCO datasets which identify different objects from captured images. Finally the voice feedback is given to the blind people through Bluetooth earphone. The proposed system uses python as programming language with raspberry pi operating system. The device will help them to easily identify different objects from surrounding.**

Keywords—Mobile Net SSD algorithm, Raspberry-Pi, Camera module, Bluetooth Earphone, Pascal VOC and COCO datasets.

I. INTRODUCTION

In this world millions of people are suffering from vision impairment, vision is the most important sense of human being as our daily activities are rely on vision. Object detection as one of the important applications in the field of computer vision has been the focus of research, and convolution neural network has made great progress in object detection[1].Many people are suffering from blindness they face some difficulties while moving around the surrounding environment; blindness makes normal life very difficult. According to the World Health Organization, 314 million people in the world are visually impaired and out of them 15% blind. 13% of visually impaired people reside in developed countries [2].Visual impairments can highly disrupt human's normal activities as simple as recognize all the things around them properly[3]. They can sense sounds and movement but not able to see object in front of them, they move around based on their sense and experiences.Machine learning is getting popular in all industries with the main purpose of improving revenue and decreasing costs; by using Machine learning technique they automate and optimize their process to solve challenging tasks very efficiently [11].

In this project, we build a real time object detection and recognition for blind people ,this system will help the blind people to identify different object from the surrounding and will give audio feedback to the user. In this system we use Mobile Net SSD (Single Shot detector) algorithm which **takes only one shot to detect multiple objects present in an image**. In Section 2 discuss different components used in proposed system with Mobile-Net SSD algorithm. The testing and result discussions are in Section 3. Then the report concluded with Section 5.

II. PROPOSED SYSTEM

The proposal presents object detection and identification device for visually impaired people. In this project, we have proposed a device for object identification with voice feedback for the visually challenged. The proposed fully integrated system has a camera as an input device to feed the real time images of object surrounded by blind people and object detection and identification is done by Mobile-Net SSD algorithm. The proposed system generally exploit a single camera to capture images of the scene in front of the user[4]. A methodology is implemented to the recognition of different objects and then provides voice feedback. As part of the software development, the Open CV (Open source Computer Vision) libraries are utilized to capture image of object. Using Mobile-Net SSD algorithm object is detected from captured images. The device is a voice enabled system that would direct the visually challenged

person in their day to day works [5]. The proposed system is trained with Pascal and COCO datasets to identify objects from images. After successful recognition, text is converted into audio output[6].

A. System Design

The system design phase provides solution to the matter specified by requirement document. System design which is additionally called as high level design, aims to spot the various modules that ought to be present within the system, and the way these different modules interact with one another to provide the specified outputs.

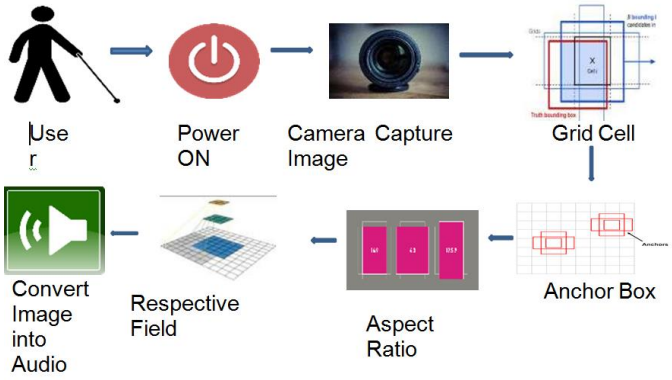


Fig 1.1 System Architecture

B. System Methodology

Fig 1.1 shows the system architecture of proposed system. User first has to start the device for further processing. The device will capture the real time images from surrounding. Then MobileNet SSD algorithm will apply on the images. SSD first divides the input image using grid cell. This grid cells are responsible for detecting object in that region of image. Grid cells are assigned with multiple anchor boxes which are responsible for size and shape within a grid cell. The ratio parameter can be used to specify the different aspect ratio of anchor box associates with each grid cell at each zoom level. The detected image is identified with Pascal and COCO datasets. Finally, the identified object is sent as an audio output to the user.

C. System Implementation

1] Open CV python

OpenCV is an open-source library for computer vision, machine learning, and image processing. OpenCV supports a range of programming languages like Python, C++, Java, etc. This library has ability to process different images and videos to spot objects, faces, and even the handwriting of a personality's. OpenCv is a library of Python bindings designed to resolve computer vision problems.

2] Raspberry Pi 3(Model b+)

The Raspberry Pi 3 (Model b+) is that the latest product within the Raspberry Pi 3 range. The camera module come up with 64bit quad core processor running at 1.4GHz, dual-band 2.4GHz, 5GHz wireless LAN, and Bluetooth 4.2/BLE, faster Ethernet, and also PoE capability via a separate PoE HAT The dual-band wireless LAN also come up with modular compliance certification, allowing the board which to be designed into end products with significantly reduced wireless LAN compliance testing, improving cost in addition as time to promote. The Raspberry Pi 3 Model b+ maintains the identical mechanical footprint as maintain by both the Raspberry Pi 2 Model B and also the Raspberry Pi 3 Model B.

Fig. 1.2 Raspberry Pi 3 Model B+

3] Raspberry-Pi Camera module:

The camera module of Raspberry-pi is accustomed take high-definition video, also as stills photographs. Numerous people using this model for various purposes like time-lapse, slow-motion and other video cleverness. The Raspberry-Pi camera module is incredibly popular in home security applications, and also in wildlife camera traps. The camera consists of a little circuit board of size 25mm by 20mm by 9mm, which connects to the Raspberry Pi's Camera Serial Interface (CSI) bus connector via a versatile ribbon cable. The image sensor contains a native resolution of 5 megapixels and encompasses a fixed focus lens. The software for the camera module supports full resolution still images up to 2592x1944 and video resolutions of 1080p30, 720p60 and 640x480p60/90.

Fig. 1.3 Raspberry Pi Camera

4] Mobile Net SSD Algorithm:

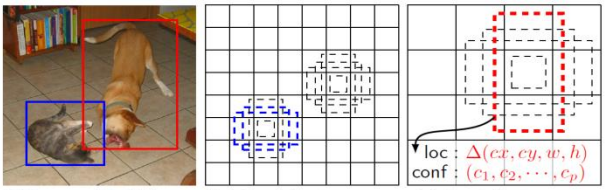
The proposed system includes Mobile Net SSD algorithm for object detection and identification. The real time images are capture by camera. Then input images are given to mobile net SSD (Single Shot Detector) algorithm which is able to detect object from images and identify those object with train dataset.

The Mobile-net SSD (Single Shot Detector) model is a Single-Shot multibox Detection network intended to perform object detection. The model is implemented using the Caffe framework. MobileNet SSD algorithm is developed by Google researcher teams. It's developed to keep up the balance between the 2 object detection methods which are YOLO and RCNN.

SSD is quicker as comparing to RCNN. In R-CNN we need two shots one for generating region proposals and one for detecting objects within frame whereas in SSD It can be done in a single shot.

Mobile Net SSD Algorithm:

1. **MultiBox Detector:**



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Fig.1.4 SSD: Multiple Bounding Boxes for loc and conf

- After going through the convolutions for feature extraction, we obtain a feature layer which is of size *m*×*n* (number of locations) with *p* channels, like 8×8 or 4×4 above. The 3×3 convolution is applied on *m*×*n*×*p* feature layers.
- We got K bounding boxes for every location. These K bounding boxes are present with different sizes moreover aspect ratios. The concept is, maybe for human vertical rectangle is fit, and for car horizontal rectangle is fit.
- We will compute class scores (c) and 4 offsets for every bounding box, relative to that of original default bounding box shape.
- Hence we got (*c*+4)*kmn* outputs.

2. **SSD(single shot multibox) Network Architecture:**

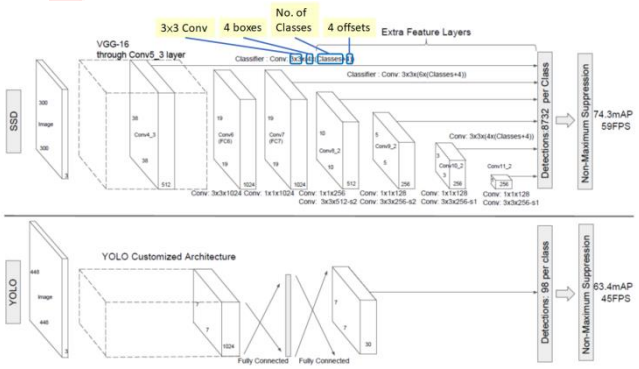


Fig. 1.5 SSD (Top) vs YOLO (Bottom)

Different layers of feature maps are also going through a small 3×3 convolution for object detection as shown above to own more accurate detection.

- Suppose, at Conv4_3, it's of size 38x38x512. Here we applied 3x3 convolutions. There are 4 bounding boxes and every will have (classes+4)outputs. Therefore, at Conv4_3, the output is 38x38x4x(*c*+4). Say, there are 20 object classes with one background class, the output is 38x38x4x(21+4) = 144,400. There's 38x38x4 = 5776 bounding boxes, in terms of total number of bounding boxes.
- Similarly for other convolution layers:
- Conv7: 19×19×6 = 2166 boxes (for each location6 boxes)
- Conv8_2: 10×10×6 = 600 boxes (for each location6 boxes)
- Conv9_2: 5×5×6 = 150 (for each location6 boxes)
- Conv10_2: 3×3×4 = 36 boxes (for each location 4 boxes)
- Conv11_2: 1×1×4 = 4 boxes (for each location 4 boxes)
- If we sum up them we got 5776 + 2166 + 600 + 150 + 36 +4 = 8732 boxes in total. There are total 7×7 locations at the end with 2 bounding boxes for every of those locations in YOLO algorithm. As YOLO only got 7x7x2 =98boxes. Therefore, SSD has 8732 bounding boxes which are way more than that of YOLO.

1. **Loss Function:**

$$L(x, c, l, g) = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, l, g))$$

The loss function is principally consists of Lconf and Lloc where N is the matched default boxes.

$$L_{loc}(x, l, g) = \sum_{i \in Pos}^{N} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^{k} \text{smooth}_{\text{L1}}(l_i^m - \hat{g}_j^m)$$

$$\hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx})/d_i^w \qquad \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h$$

$$\hat{g}_j^w = \log\left(\frac{g_j^w}{d_i^w}\right) \qquad \hat{g}_j^h = \log\left(\frac{g_j^h}{d_i^h}\right)$$

Lloc is the localization loss which is the smooth L1 loss between the two parameters which are predicted box (l) and the ground-truth box (g). These parameters include the offsets for the middle point (cx, cy), width (w) likewise for height (h) of the bounding box. So this loss is analogous to the Faster RCNN algorithm.

$$L_{conf}(x, c) = -\sum_{i \in Pos}^{N} x_{ij}^{p} log(\hat{c}_i^p) - \sum_{i \in Neg} log(\hat{c}_i^0) \quad \text{where} \quad \hat{c}_i^p = \frac{\exp(c_i^p)}{\sum_p \exp(c_i^p)}$$

Confidence Loss

Lconf stands for confidence loss which is nothing but the softmax loss over multiple classes confidences (c). (by cross validation α is set to 1.) *xij^p* = {1,0}, this means that it should match *i*-th default box to the *j*-th ground truth box of category *p*.

1. **Scales and Aspect Ratios of Default Boxes:**

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m - 1}(k - 1), \quad k \in [1, m]$$

Scale of Default Boxes

Suppose there are *m* feature maps for prediction, then we are able to calculate *Sk* for the *k*-th feature map. Smin is 0.2, Smax is 0.9. So it means the scale at the lowest layer is 0.2 and the scale at the highest layer is 0.9. All the layers in between that are regularly spaced.

For each of the scale, sk, we've 5 non-square aspect ratios:

$$a_r \in \{1, 2, 3, \tfrac{1}{2}, \tfrac{1}{3}\} \quad (w_k^a = s_k\sqrt{a_r}) \quad (h_k^a = s_k/\sqrt{a_r})$$

5 Non-Square Bounding Boxes

For aspect ratio of 1:1, we got sk':

$$s'_k = \sqrt{s_k s_{k+1}}$$

1 Square Bounding Box

So, with different aspect ratios there are at the most 6 bounding boxes we will have in total.
Main Objective of SSD algorithm to detect various objects in real time video sequence and track them in real time [8].So this is often how Mobile Net SSD algorithm work on input images for object detection and identification.

C]Speech Synthesis

Speech synthesis is artificial production of human speech. A computer system that's used for speech synthesis purpose is termed a speech computer or speech synthesizer, and can be implemented in software or hardware products. A text-to-speech (TTS) conversion system converts normal language text into audio form speech.

> Synthesized speech can be created by merging pieces of recorded speech that are stored in an exceedingly database. A synthesizer can build a model of the vocal tract and other human voice characteristics to form a totally "synthetic" voice output. Voice output works through TTS (text to speech)[7].

The quality of performance of a speech synthesizer is measured by its similarity to the human voice and by the ability to be understood clearly. A perfect text-to-speech system allows those that are with visual impairments or reading disabilities to listen to written words on a home computer.

A text-to-speech synthesizer system is combination of two parts front end and back end. The front-end performs two major tasks. At Starting, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words. This process performed by the front end is termed as text normalization, pre-processing, or tokenization. The front-end of TTS system then assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The method of assigning phonetic transcriptions to words is named as text-to-phoneme conversion. Phonetic transcriptions and prosody information together form the symbolic linguistic representation that is output by the front-end. The back-end is also referred to as the synthesizer—then converts the symbolic linguistic representation into sound. In some systems, this part includes the computation of the target prosody (pitch contour, phoneme durations), which is then imposed on the output speech.

Speech synthesis using python

There are different APIs available to convert text to speech in Python. One among such APIs is that the Google Text to Speech API commonly referred to as the gTTS API. gTTS is a easy to use tool which converts the text entered, into audio form which can be saved as a mp3 file.

The gTTS API supports many languages including English, Hindi, Tamil, French, German and plenty of more. The speech output can be delivered in any one of the two available audio speeds, which are fast or slow. However, as of the most recent update, it's impossible to alter the voice of the generated audio.

We have used gTTS for converting text output into audio form in order that it'll be helpful to the blind people.

III. Result and Analysis

As mentioned earlier we've got used SSD algorithm for object detection and identification, there are two types of SSD algorithm, is available,

SSD300:-300×300 input image, lower resolution, faster.
SSD512:-512×512 input image, higher resolution, more accurate.

We have used SSD300 for our project. If we compare SSD algorithm with YOLO algorithm, For YOLO, detection is a straightforward regression dilemma which take an input image and learns the class possibilities with bounding box the coordinates. YOLO divides each and each image into a grid of S x S and each grid predict N bounding boxes and confidence. The confidence value reflects the precision of the bounding box and whether the bounding box in point of fact contains an object in spite of the defined class. YOLO algorithm even forecasts the classification score for each box for every class. You'll be able to even merge both the classes to figure out the prospect of each class being present in a predicted box.

So, all the overall SxSxN boxes are forecasted. On the opposite side, most of those boxes have lower confidence scores and if we set a doorstep say 30% confidence, we will get eliminate most of them.

SSD includes a better balance between swiftness and precision.SSD runs on a convolutional network input image only 1 time and computes a feature map. A Single Shot Multi-Box Detector (SSD) is an object detection approach in images using a deep neural network [9] Now, we run small 3×3 sized convolutional kernel on this feature map to foresee the bounding boxes and categorization probability.

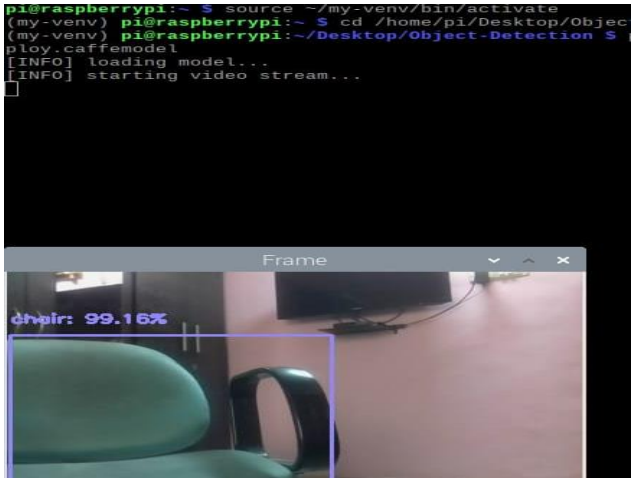
SSD uses anchor boxes at a variety of aspect ratio comparable to Faster-RCNN and learns the off-set to a specific extent than learning the box. So as to hold the scale, SSD algorithm predicts bounding boxes after multiple convolutional layers. Since each convolutional layer functions at a various scale, it's able to detect objects of a mix scales. A single neural network predicts bounding

boxes and class probabilities directly from full images in one evaluation[10].

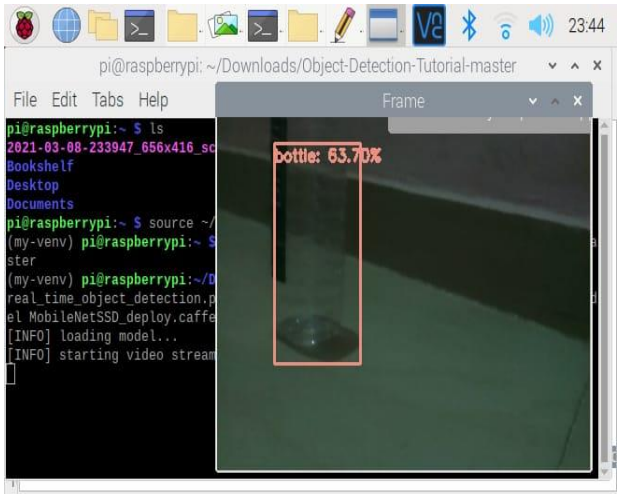
SSD is pretrained on prototxt model and caffemodel. A prototxt is a text file that holds information about the structure of the neural network: A listing of layers within the neural network. The parameters of every layer, like as its name, its type, input dimensions, and output dimensions, all the connections between the layers. Caffe model is deep learning model which carries with it COCO dataset and PASCAL VOC dataset which contains over 1400 images for traning and testing of object detection and identification models.

In our project we've got tested our model on around 50 images, out of which around 44 images have identified by the model correctly.

Some of the results are as follows:



Here, In this image chair is detected 99.16%

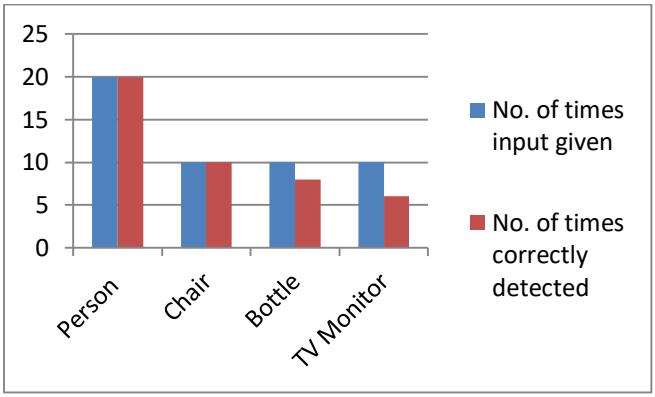


In this image bottle is detected 63.70%. If confidence value is smaller amount than 20 % then it will be considered as background by the model.

Results:

Name of Input image	No. of times input given	No. of times correctly detected	Correctness accuracy (%)
Person	20	20	100%
Chair	10	10	100%
Bottle	10	8	80%
TV Monitor	10	6	60%

Analysis of results:



IV. Conclusion

We present the system for blind people based totally upon object detection and identification. This system makes use of SSD (Single Shot Detection) to identify objects. Objects detection is used to find objects in the real world from images of the world. The raspberry-pi camera is used to capture real time images from surrounding. To provide voice feedback, the identified image is converted into audio format using GTTS (Google text to speech) module. So, the proposed system will assist the blind people to discover different objects from surrounding and could deliver sound as output to the user. This system is used in actual time object detection. The navigation system is highly-priced which isn't always affordable to blind people. So, this project goal is to help blind peoples.

.

V. Future Scope

Future scope can be, identifying user's known people when they passed by them. Also to compute the gap between the blind person and every object, on the way to effortlessly recognize how lengthy item from them. To make lifestyles extra simpler of blind human beings the night vision mode will be available in inbuilt camera.

VI. References

[1]Xinyi Zhou, Wei Gong, WenLong Fu, FengtongDu,Application of Deep Learning in Object Detection, 2Information Engineering School, Communication University of China,2017

[2]VikkyMohane,Prof. Chetan Gode,Object Recognition for Blind people Using Portable Camera,2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (WCFTR'16).

[3]Joe Yuan Mambu, Gerent Keyeh ,Elisa Anderson ,FakultasIlmu, Billy Dajoh,Blind Reader: An Object Identification Mobilebased Application for the Blind using Augmented Reality Detection, 2019 1st International Conference on Cybernetics and Intelligent System (ICORIS).

[4]HanenJabnoun, FaouziBenzarti,and Hamid Amiri, Object recognition for blind people based on features extraction, IEEE IPAS'14: INTERNATIONAL IMAGE PROCESSING APPLICATIONS AND SYSTEMS CONFERENCE 2014

[5]Joe Louis Paul I, Sasirekha S ,Moohana Priya p,Mohanavalli S ,Smart Eye for Visually Impaired-An aid to help the blind people, Second International Conference on Computational Intelligence in Data Science (ICCIDS-2019)

[6]Wang Zhiqiang1 , Liu Jun ,A Review of Object Detection Based on Convolutional Neural Network, Proceedings of the 36th Chinese Control Conference July 26-28, 2017, Dalian, China

[7] Akhila.S, DishaMRani, Divyashree.D, Varshini.S.S,Smart Stick for Blind using Raspberry Pi,International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181,ICACT - 2016 Conference Proceedings

[8]Chandan G, Ayush Jain, Harsh Jain, Mohana , Real Time Object Detection and Tracking Using Deep Learning and OpenCV, IEEE Xplore Compliant Part Number:CFP18N67-ART; ISBN:978-1-5386-2456-2

[9]Reagan L. Galvez, Argel A. Bandala, Elmer P. Dadios, Object Detection Using Convolutional Neural Networks, Proceedings of TENCON 2018 - 2018 IEEE Region 10 Conference (Jeju, Korea, 28-31 October 2018).

[10]Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi ,You Only Look Once: Unified, Real-Time Object Detection, 2016 IEEE Conference on Computer Vision and Pattern Recognition

[11] Patil, V., Ingle, D.R. An association between fingerprint patterns with blood group and lifestyle based diseases: a review. ArtifIntell Rev 54, 1803–1839 (2021). https://doi.org/10.1007/s10462-020-09891-w