GOCE DELCEV UNIVERSITY - STIP
FACULTY OF COMPUTER SCIENCE

ISSN 2545-4803 on line

BALKAN JOURNAL OF APPLIED MATHEMATICS AND INFORMATICS (BJAMI)

YEAR 2021

VOLUME IV, Number 2

GOCE DELCEV UNIVERSITY - STIP
FACULTY OF COMPUTER SCIENCE

ISSN 2545-4803 on line

BALKAN JOURNAL OF APPLIED MATHEMATICS AND INFORMATICS

(BJAMI)

YEAR 2021 VOLUME IV, Number 2

AIMS AND SCOPE:
BJAMI publishes original research articles in the areas of applied mathematics and informatics.

Topics:
1. Computer science;
2. Computer and software engineering;
3. Information technology;
4. Computer security;
5. Electrical engineering;
6. Telecommunication;
7. Mathematics and its applications;
8. Articles of interdisciplinary of computer and information sciences with education, economics, environmental, health, and engineering.

Address of the editorial office
Goce Delcev University – Stip
Faculty of philology
Krste Misirkov 10-A
PO box 201, 2000 Štip,
Republic of North Macedonia

BALKAN JOURNAL
OF APPLIED MATHEMATICS AND INFORMATICS (BJAMI), Vol 3

ISSN 2545-4803 on line
Vol. 4, No. 1, Year 2021

EDITORIAL BOARD

CONTENT

UDC: 004.89.032.26:004.932’1

AN OVERVIEW OF IMAGE RECOGNITION AND REAL-TIME OBJECT DETECTION

MIRJANA KOCALEVA AND SASO KOCESKI

Abstract. Nowadays neural networks, deep learning and computer vision give us the best solutions for many problems in Artificial intelligence such as image recognition, speech recognition, natural language processing, price and load forecasting, healthcare, marketing, recommendation systems, etc. This paper presents an overview of research in the field of image recognition done in recent years. First an outline of object recognition and deep learning methodologies are given. Then convolutional neural networks as architecture for deep learning are explained. In addition, image annotation techniques and tools are evaluated, to choose those with the best performance and accuracy. At the end, the frameworks for real-time object detection in prospect of our research are discussed.

1. Introduction

The growing popularity of the Artificial Intelligence (AI) and its application in various fields, starting from tourism through medicine [4], biology, education, robotics [1] and in economy is mainly due to the apparatus i.e., the models and techniques used to mimic human reasoning, to learn and improve during time. A process or technology in the range of computer vision and Artificial Intelligence (AI) with the main task of identifying objects (things, entities) in a static image or in a dynamic digital video can be defined as object recognition. Humans can perceive many objects in images, although the images are sometimes in different sizes, rotated or translated, even when the images are seen from different points of view. Although this task is easy for humans, it is still a challenge for the systems for computer vision [8], [9].

The algorithms for object recognition are using techniques established on appearance or character [8]. According to [9], the methods used for object identification include 3D models, component identification, edge detection, and analysis of appearances from different angles. Today there are many models used for object recognition. Some of them are as follows [8], [9]:

- Deep learning models like CNNs
- Feature extraction and machine learning models
- Models such as maximally stable external regions and speeded up robust features
- Some matching approaches
- The Viola-Jones algorithm

- Template matching
- Colour based
- Shape based
- Active and Passive
- Division of images and examination of image regions.

Authors in [10] show that with the development of Convolutional Neural Network (CNN) architectures, a computer can be better than a human in the object recognition task under some specific circumstances (as in the case of face recognition). Therefore, the focus of this paper is on Deep learning and Convolutional neural networks.

Until 2006, it was hard to use some new methods for training neural networks. Therefore, in 2006 a new technique for deep learning in artificial neural networks (ANNs) was discovered (Figure 1). Now these techniques are known as deep learning (hierarchical learning) [11]. However, what exactly is deep learning? Deep learning can be defined as a machine learning technique that learns features and tasks of ANNs that contain not only one hidden layer directly from data. All images, texts, or sounds can represent the data. There are 3 types of learning: supervised, partially supervised, and unsupervised [11]. Unsupervised learning is useful when we want to explore data but do not yet have a specific goal or we do not know at all the type of information included in the data. For unsupervised learning, the most used technique is cluster analysis [8]. The algorithms for supervised learning know what type of data they will receive for processing. In addition, this type of algorithms knows the type of data that they will give us as a result. They also train the current model for generating predictions for the new data they will receive. Classification and regression are the most used forms for supervised learning [8]. Deep Learning is almost everywhere: natural language processing, object classification, music and arts, object detection, social network filtering, bioinformatics, segmentation, pose estimation, image captioning, question answering, machine translation, speech recognition, audio recognition, and robotics [11].

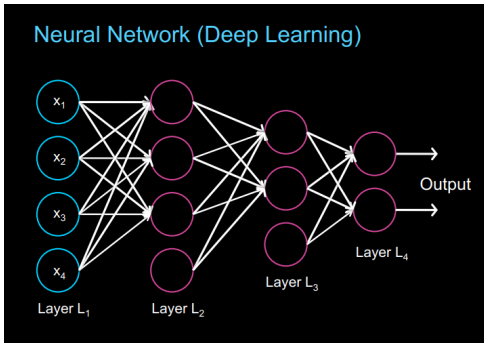

Figure 1. Deep learning network [11]

Convolutional neural networks (CNNs) are a type of artificial neural network for deep learning. CNNs architecture is given in Figure 2. On the other hand, we can also define CNNs as several layers of convolutions with nonlinear activation functions applied to the results. In a traditional neural network, we used a fully connected layer where we related respectively each input neuron to a neuron in the output. Nevertheless, when we operate with CNNs we do not do that. In that case, we simply use convolutions over the input layer to compute the output, and we have a relationship between each region of input neurons with one output neuron. This relationship is known as local connection. Each layer applies hundreds or thousands of different filters to combine the results. There are also pooling layers or downsampling layers. From several options in this layer category, maxpooling is the one that is most used [12].

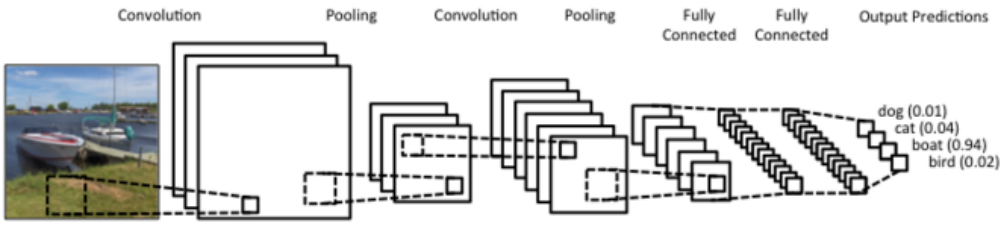

Figure 2. CNN architecture [12]

2. Image annotation

We often think that when we see something it is without investing a lot of effort. However, the visual system is sometimes very slow, and we think we have captured all the parts of the image, but we have not. Image annotation means precisely to understand what is in the image and using metadata that is being added to images to facilitate access to certain image search [2].

There are three techniques for image annotation: manual, semi-automatic and automatic. We are talking about manual annotation when users have to write a description with few keywords when reviewing images. Manual image annotation takes a lot of time. Also, it is an expensive, hard and boring process [2]. The automated image annotation (AIA) system automatically assigns a set of keywords that help to understand the image semantic content using a computer system. One of the biggest advantages of the automatic image annotation is the efficiency, but not the accuracy [5]. The semi-automatic image annotation combines manual annotation efficiency and automatic annotation accuracy. In this type of annotation, users are encouraged to provide an initial query and feedback while browsing images [3].

Today there are many different tools for manual, automatic and semi-automatic image annotation. Some of them that we considered are given in Table 1.

Table 1. Tools for image annotation

Annotation	Tool	Features	Format
Manual	LabelMe [13]	-WEB-based image annotation tool -Allows researchers to label images and share the annotations with the world	XML format
	BBox-label-tool [14]	-A tool for labelling objects in images with drawing bounding boxes around the objects	Text file
Semi-automatic	Ratsnake [15]	-Fast segmentation and annotation of images with polygons, grids or both -Ontology-based image annotation -Annotation of web content -Image annotations for the semantic web -Publication of annotations on the web -Transformation of binary image masks to polygon annotations -Compatibility with the LabelMe annotation tool	custom text, owl, LabelMe XML
	Alp's Large Image Annotation Tool (LIANT) [21]	-Works for .jpg and .png files only -It can read label box data from .txt files in the same directory. -Allows to create label data from scratch only -Zoom in/out -Rotation on image and on label	.txt file
Automatic	LEAR [26]	-Pixel-wise object annotation -Zoom in/out -Different brush sizes (circle shape) -Line drawing -Flood filling -Different colour types: background, object, occluded object -Different drawing modes: over all colour types or only over a specific colour type (i.e., masked) - A mask file (in .png format) is created for each object separately	.bmp format
	Annotation and Image Markup – AIM [27]	-The goal of this tool is to develop a mechanism for modelling, capturing, and serializing image annotation and mark-up data that can be adopted as a standard by the medical imaging community	No standard format

From the software given in Table 1., the most used in scientific research are BBOx [16], [17] and LabelMe [18], [19]. In papers [16] and [17], the Bbox-label-tool for manual labelling of images in two different fields like hospitals and senior homes and transport is given. Matija et al. [16] present the procedure that they used for training of CNNs on a group of aerial images. The results obtained from object recognition show that CNN can detect and classify objects with around 97.5% accuracy (high accuracy). Furthermore, with the YOLO platform, they have tracked, detected, and classified objects from video feeds provided by unmanned aerial vehicles in real-time. In addition, in [17] Jin et al. discuss the rapid development of technology and how Web has become the largest

encyclopaedic database. However, they also said that users using a web browser could only get some surface information, because it is hard to get deep web information. For extraction data in the deep web, they must use some methods. So, for that reason, a method for data region locating based on CNN and a segmentation algorithm are proposed.

Image annotation is labour-intensive work when we must work manually. On the other hand, manual annotation is the best and the most accurate way for image annotation used. One way to overcome this problem is with crowdsourcing. Therefore, Phani et al. [18] mainly discuss crowdsourcing. They also talk about how advances in digital photography resulted in millions of images uploaded on the Web. They used LabelMe image annotation datasets for their research. They choose LabelMe for construction of their own dataset because it is faster and cheaper than the traditional methods. In his paper Phani also talks about human-computer interaction and some issues correlated with images.

Another problem in the process of annotation is a large scene collection and 3D coordinates extraction. Paper [19] presents that problem. There the LabelMe annotation tool and its annotation corpus are described. The main aim of LabelMe was to support image database with great number of objects that are annotated. A 2-D semantic layout visualization was performed with the help of the large scenes collection and labelling and it was shown how to extract a 3-D coordinates of images in different scenes using only the object annotations provided by users.

Image annotation tools can be standalone (also known as thick client architecture) or web based. The main difference between them is that standalone annotation tools are running on a desktop environment and web-based ones are running on a web. The standalone annotation tool is not so good because it depends on third party controls and platform; the process of client's update is long and there is also a problem when many clients are connected to the server at the same time. The advantage is that this kind of architecture can work even offline. Web based annotation tools have less functionality because they need a stable internet connection and an adequate web browser. On the other hand, this solution is good because it is centralized, easy for update and deployment and it is independent of a platform [6]. From the papers cited above, we can conclude that tools for image annotation (BBox and LabelMe) are used in many areas of research, such as hospitals and senior homes, transportation, web browsers, crowdsourcing, human computer interaction, surgery, real video, etc.

3. Real-Time Object Detection and applications

There are many problems in the scope of computer vision. One of the main problems is the detection of an object or the process of identifying the items in an image and to discover where the place of that item is in the image [20]. This process is hard because we cannot always see the image clearly and from the correct viewpoint. From here comes the need for using different frameworks for object detection in real-time. One framework is YOLO, and the use of YOLO is presented respectively in [23], [24], [25]. Beside YOLO, there are many other frameworks such as Faster R-CNN [22], generative

framework [28], robust object tracking [29], Viola–Jones object detection framework [7] (Table 2).

Table 2 Different frameworks for object detection

Framework	Features				
	Neural network type	Backbone Feature Extractor	Speed	Anchor Box	Detection precision
YOLO	Fully convolutional	Darknet	Faster than R-CNN	K-means from coco and VOC, 9 anchors boxes with different sizes	low
Faster R-CNN	Fully convolutional	VGG-16 or other feature extractors	Slower than YOLO	9 default boxes with different scales and aspect ratios	high
Viola–Jones object detection	No convolutional	/	2 frames per second	more than one rectangular feature	high

YOLO is one of the most famous frameworks for object detection. YOLO presents a neural network for performing object detection in real-time. In addition, the framework uses only one convolutional neural network for the whole image for object detection using bounding boxes. These bounding boxes have a weight that comes from the probabilities of prediction. The basic YOLO model operates in about 45 frames per second and the Fast YOLO (a smaller version of the model) operates in about 155 frames per second. YOLO is a good network because it learns a base representation of objects [22].

The architecture of YOLO is given with the following image:

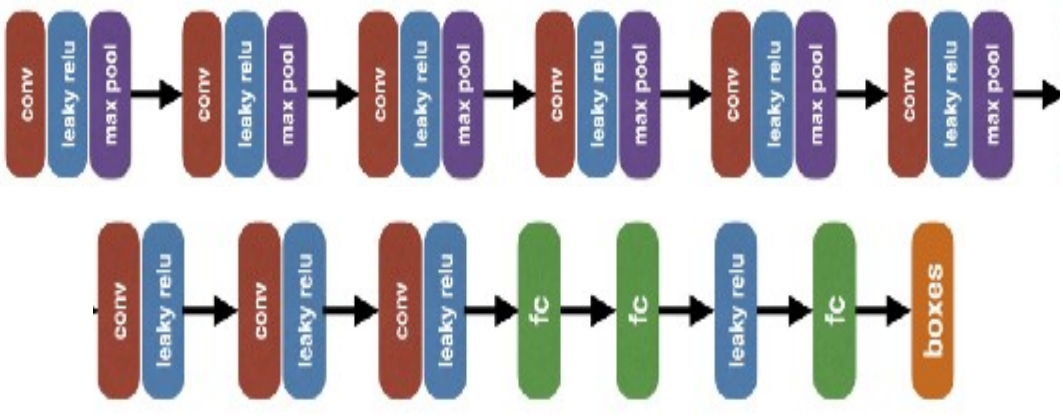

Figure 3. YOLO architecture [22]

We use the YOLO framework for solving different problems in object detection such as using one controller for navigating around the scene. This problem exists when we have a question answering system [23] that answers questions that used an autonomous agent for communication with a visual environment. The agent must navigate around the scene, gain visual understanding of the scene elements, interact with objects, and plan for a series of actions conditioned on the question. For that reason, they use a group of controllers instead of one controller.

The other problem is tracking location. The CCN [24] used the history of locations as well as the distinctive visual features learned by the deep neural networks. They propose to link together visual features produced by the CNN with the information about regions. For forecasting tracking locations, they used backsliding. From the result they conclude that their tracker is competitive because it offers low computational cost. Here Guanghan et al. proposed a new method named ROLO as extension of YOLO.

Another problem appears with training video frames. This problem is considered in [25] where the authors train a pseudo-labeller first. The pseudo-labeller is first trained individually on the video frames that are labelled and then subsequently applied to all frames. After that, they trained a recurrent neural network for intake sequences of the pseudo labelled frames and optimized an objective that encourages both the efficiency on the destination frame and the flexibility over continuous frames. It has been experimentally proven that adjacent frames could provide some important information, even if some labels are missing.

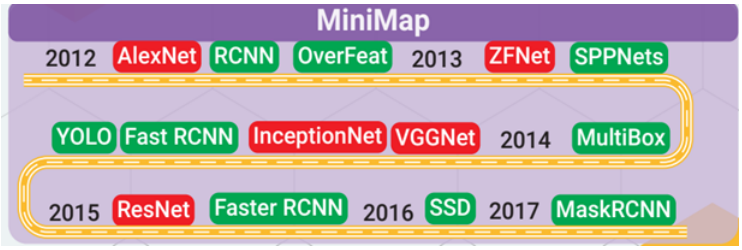

Figure 4. Modern history of Object Recognition Infographic [25]

Fasel et al. [28] create a model for defining the probability of image generation and develop an algorithm for object finding and characteristics within that framework. With the model, images are presented as a collage, and this model is tested for finding faces and eyes in images using different lights.

Alan J. L. et al. [3] and Fatih Porikli [29] talk about video object detection. Paper [3] describes a method for extracting moving targets detected using the pixel wise difference between continuous image frames from a video and [29] presents a survey of methods for tracking and object detection on video (camera). In [3] images are classified in three categories, namely: human, vehicle or background clutter. Once classified, the targets are followed by a pair of temporal anticipation and pattern matching. The developed system recognizes targets of interest and continually tracks them over big distances and duration

of time despite obstruction. In [29] authors also make a contrasting of image complexities, which must be on minimum. Those methods are designed to be performed in many difficult conditions such as unpredictable motion, lighting change, and noise adulteration.

In the last paper [7], an algorithm for text detection in city pictures was designed and the authors wanted to reduce the criterion for creating cascade that cares about time needed for tests and time needed for pre-processing. Their algorithm runs 40 frames per second. This speeds up factor results in creating two systems, Smart Telescope and Signfinder, which use text detection to design systems to help the blind and visually impaired. According to the papers above, we can see that Real-Time Object Detection has applications in many different fields. Some of them are Interactive Question Answering system, visual object tracking, system for the blind in real-time, railway (for defection of the fasteners), face detection, videos, images, autonomous driving, robots, vehicles, text detection, real-time system for visually impaired people and so on.

4. Conclusion

Object recognition and image annotation are some of the most important techniques in the field of computer vision. In this article, we make an overview of recent studies in which researchers used different tools for image annotation and frameworks for Real-Time Object Detection. Our next aim is to collect images and annotate them, and then to use some corresponding framework for object detection. Considering the reported findings, we are planning to choose the YOLO framework for our object recognition tasks, because compared with other frameworks, it is faster and works with only one CNN network. From the three types of annotation, we are going to choose the manual image annotation because it is more precise and effective compared to other methods, even though it is slower. In addition, respectively, from the spectrum of tools for manual annotation, we will choose BBox label tool for labelling our own image database. Moreover, this tool exports data in .txt files and therefore we can easily convert these files into YOLO format files.

References

[1] Koceska, Natasa, Saso Koceski, Francesco Durante, Pierluigi Beomonte Zobel, and Terenziano Raparelli. "Control architecture of a 10 DOF lower limbs exoskeleton for gait rehabilitation." International Journal of Advanced Robotic Systems 10, no. 1 (2013): 68.

[2] Sumathi, T., Devasena, C. L., & Hemalatha, M. (2011). An overview of automated image annotation approaches. International journal of research and reviews in information sciences, 1(1), 1-5.

[3] Lipton, A. J., Fujiyoshi, H., & Patil, R. S. (1998, October). Moving target classification and tracking from real-time video. In Applications of Computer Vision, 1998. WACV'98. Proceedings. Fourth IEEE Workshop on (pp. 8-14). IEEE.

[4] Koceski, Saso, and Natasa Koceska. "Evaluation of an assistive telepresence robot for elderly healthcare." Journal of medical systems 40, no. 5 (2016): 121.

[5] Sami, M., El-Bendary, N., & Hassanien, A. E. (2012, October). Automatic image annotation via incorporating Naive Bayes with particle swarm optimization. In Information and Communication Technologies (WICT), 2012 World Congress on (pp. 790-794). IEEE.

[6] Molin, P., & Löfberg, M. (2005). Web vs. Standalone Application.

[7] Chen, X., & Yuille, A. L. (2005, June). A time-efficient cascade for real-time object detection: With applications for the visually impaired. In Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on (pp. 28-28). IEEE.
[8] MathWorks, Object recognition, Machine learning. MathWorks Inc. South Natick MA.
[9] Khurana, K., & Awasthi, R. (2013). Techniques for object recognition in images and multi-object detection. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2(4), pp-1383.
[10] Hien, D. H. T. (2017). Modern History of Object Recognition—Infographic.
[11] Gavves, E. (2016). Introduction to Neural Networks and Deep Learning. UvA Deep Learning Course.
[12] Britz, D. (2015). Understanding Convolutional neural networks for NLP. URL: http://www. wildml. com/2015/11/understanding-convolutional-neuralnetworks-for-nlp/(visited on 21/02/2018).
[13] LabelMe http://labelme.csail.mit.edu/Release3.0/(visited on 11/01/2018).
[14] Bbox https://github.com/puzzledqs/BBox-Label-Tool (visited on 11/01/2018).
[15] Ratsnake https://is-innovation.eu/ratsnake/(visited on 21/03/2018).
[16] Radovic, M., Adarkwa, O., & Wang, Q. (2017). Object Recognition in Aerial Images Using Convolutional Neural Networks. Journal of Imaging, 3(2), 21.
[17] Liu, J., Lin, L., Cai, Z., Wang, J., & Kim, H. J. (2017). Deep web data extraction based on visual information processing. Journal of Ambient Intelligence and Humanized Computing, 1-11.
[18] Kidambi, P., & Narayanan, S. (2008, July). A human computer integrated approach for content- based image retrieval. In Proceedings of the 12th WSEAS International Conference on Computers, Recent Advances in Computer Engineering (pp. 691-696).
[19] Torralba, A., Russell, B. C., & Yuen, J. (2010). Labelme: Online image annotation and applications. Proceedings of the IEEE, 98(8), 1467-1484.
[20] Amit, Y., & Felzenszwalb, P. (2014). Object Detection. Computer Vision: A Reference Guide, 537-542.
[21] Alp's Large Image ANnotation Tool (LIANT) https://alpslabel.wordpress.com/2017/04/09/alps-large-image-annotation-tools-liant-for-detectnet//(visited on 05/06/2018)
[22] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).
[23] Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., & Farhadi, A. (2017). IQA: Visual Question Answering in Interactive Environments. arXiv preprint arXiv:1712.03316.
[24] Ning, G., Zhang, Z., Huang, C., Ren, X., Wang, H., Cai, C., & He, Z. (2017, May). Spatially supervised recurrent convolutional neural networks for visual object tracking. In Circuits and Systems (ISCAS), 2017 IEEE International Symposium on (pp. 1-4). IEEE.
[25] Tripathi, S., Lipton, Z. C., Belongie, S., & Nguyen, T. (2016). Context matters: Refining object detection in video with recurrent neural networks. arXiv preprint arXiv:1607.04648.
[26] LEAR https://lear.inrialpes.fr/people/klaeser/software_image_annotation (visited on 05/06/2018)
[27] Annotation and Image Markup – AIM https://wiki.nci.nih.gov/display/AIM/Annotation+and+Image+Markup+-+AIM(visited on 05/06/2018)
[28] Fasel, I., Fortenberry, B., & Movellan, J. (2005). A generative framework for real time object detection and classification. Computer Vision and Image Understanding, 98(1), 182-210.
[29] Porikli, F. (2006). Achieving real-time object detection and tracking under extreme conditions. Journal of Real-Time Image Processing, 1(1), 33-40.

Mirjana Kocaleva
Goce Delcev University of Stip,
Faculty of Computer Science
North Macedonia
mirjana.kocaleva@ugd.edu.mk

Saso Koceski
Goce Delcev University of Stip,
Faculty of Computer Science
North Macedonia
saso.koceski@ugd.edu.mk