INFINITY

IMMERSE. INTERACT. INVESTIGATE.

Technical Brief

Object Detection and Recognition in Visual Content Understanding

DISSEMINATION LEVEL PUBLIC

PARTNER

CERTH

AUTHOR

Despoina Touska, Konstantinos Gkountakos, Ourania Theodosiadou, Theodora Tsikrika, Stefanos Vrochidis

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 883293. The content of this document represents the view of the authors only and is their sole responsibility. The European Commission does not accept any responsibilty for any use that may be made of the information it contains.

Object Detection and Recognition in Visual Content Understanding

1. Introduction[1]

Visual content understanding is a rapidly developing field of computer vision that focuses on enabling machines to interpret and analyse visual information, such as images and videos, in a manner similar to human perception. This includes a range of algorithms able to extract meaningful information from visual data in order to detect and recognise objects,[2] classify images,[3] segment regions of interest,[4] recognise object activities[5] and many others. With the recent advancements in machine learning and deep learning, these aforementioned tasks are typically implemented using intelligent systems, which are trained on large datasets[6] of annotated visual data.

Visual content understanding has a wide range of potential applications and can bring significant benefits to various research and professional fields. These fields include but are not limited to robotics,[7] autonomous vehicles,[8] medical imaging[9] and others. For instance, by understanding visual content, robots can learn to recognise and interpret non-verbal cues, such as facial expressions or gestures of humans, enhancing human-robot interaction. In the field of autonomous driving, visual content understanding applications are crucial for enabling self-driving cars to understand the objects in their surroundings, avoid occlusions and ensure safe navigation in their environment. In healthcare, medical imaging analyses, disease diagnosis, as well as treatment planning, are some tasks that visual content understanding tools can benefit, assisting the medical staff.

Security is another field in which visual content understanding algorithms can aid in accurately detecting suspicious behaviour.[10] Thereby, surveillance methods have gained growing attention from the research community as a measure to identify and prevent crimes. These methods employ diverse technologies to process data of different types, ranging from textual to visual evidence, to support Law Enforcement Agencies (LEAs) in their investigation processes. Visual evidence, in particular, comprises images and videos captured from surveillance cameras placed in both indoor and outdoor environments. As visual information becomes increasingly complex, such as in crowded scenes where many objects appear, automatic methods for object detection and recognition become necessary. Advanced technologies,[11] incorporating learning techniques, provide a solution to the problem by combining speed and accuracy when processing visual data, thus eliminating the need for human interference.

2. Related work

This section presents a number of state-of-the-art computer vision approaches for object detection and recognition that match the objectives of the INFINITY project. The section is divided into two subsections, the first of which discusses object detection and tracking techniques, while the second one addresses abandoned object detection techniques.

2.1 Object Detection and Tracking

The goal of an object detection system is to scan digital images, search and locate instances of each one of the Objects of Interest (OoI), as well as recognise their identity. Research has been conducted over the years to improve object detection by introducing futuristic models that incorporate Artificial Intelligence (AI) technologies to boost performance. As for their applicability, object detection algorithms serve as a basis for many other important computer vision tasks, such as autonomous driving,[12] traffic control[13] and activity recognition.[14]

Deep Learning (DL) has been the solution for object detection in recent years. It uses deep feature representations and complex workflows to boost performance. Convolutional Neural Networks (CNNs) trained on large datasets are especially effective for producing fast and accurate object predictions. Object detectors are categorised into two architectures: single-stage object detectors and two-stage object detectors. Single-stage object detectors directly classify and predict the bounding boxes of object candidates, while two-stage object detectors first extract Regions of Interest (RoI), and then classify and predict them.

YOLO is a single-stage object detector, which has become popular in industrial applications, as it provides an efficient lightweight design with high performance. YOLO involves a single neural network trained end-to-end to predict bounding boxes and class labels for each object candidate. Until today, different versions of YOLO have been launched, such as YOLOv1,[15] YOLOv2,[16] YOLOv3,[17] and YOLOv4.[18] Especially, YOLOv4 can achieve 43.5% Average Precision for the Microsoft COCO dataset at a real-time speed of approximately 65 Frames Per Second (FPS) on Tesla V100 GPU, demonstrating superior performance compared to other more complicated CNN architectures.[19]

Two-stage object detectors include Region-Based CNNs (R-CNNs),[20] Fast-RCNNs,[21] Faster-RCNNs,[22] Mask-RCNNs,[23] and others. R-CNN uses a selective search algorithm to find RoIs in the input image, and then a CNN architecture to classify each one of them.[24] Fast-RCNN proposes a RoI pooling layer to reduce the input image size, allowing faster feature extraction and classifier training compared to the original R-CNN architecture.[25] In addition, Faster-RCNN includes a Region Proposal Network (RPN), which is responsible for generating high-quality RoIs and improves the efficiency of the entire detection system.[26] Mask-RCNN[27] adds a branch to the Faster-RCNN model[28] that is used to generate segmentation masks for each detected object, providing a more detailed representation of it.

As for object tracking, relevant algorithms can be used to assign a unique identification to each of the OoI detected in a video and build upon their past detections by linking them together to form their trajectory over time. Additionally, any newly detected object can also be used to initiate new trajectories. An object tracking algorithm should be able to identify and track targets reliably, even when they are occluded, since this is a typical case in crowded surveillance scenarios. Occlusions, i.e., the blocking of a target's view by another object or person, can make it difficult to accurately track the target across multiple frames. However, an effective object tracking algorithm should be able to account for these occlusions and still track the target accurately.

Proposed architectures try to associate detections of objects in consecutive frames, taking into account different types of cues, such as position, motion and visual appearance. The Simple Online and Realtime Tracking (SORT) algorithm processes detections frame-to-frame and link them together based on position and size features.[29] More specifically, SORT propagates the position of already created tracks to the following frame with the Kalman filter, and then associates them with the new detection, measuring their overlap with the Intersection over Union (IoU). DeepSORT algorithm is an extension of SORT with superior performance, as it incorporates deep visual features creating a more informed association metric that combines both motion and appearance information.[30]

Within the context of the INFINITY project, YOLOv4 has been utilised as the object detector. YOLOv4 is widely recognised as a state-of-the-art model in the field, renowned for its exceptional speed and accuracy, making it an excellent fit for the objectives of the INFINITY project. YOLOv4 focuses on detecting specific OoI, including persons, backpacks, suitcases, and handbags. The output of YOLOv4 provides valuable bounding boxes for the detected objects, accompanied by their respective confidence scores. Regarding object tracking, the Euclidean distance has been employed as a metric to measure the distance between the detected objects in consecutive frames. To determine the future position of an object in a frame, the object with the shortest distance is selected in comparison to others in the next consecutive frame.

2.2 Abandoned Object Detection

The detection of abandoned objects constitutes a critical component in surveillance systems that aims to prevent criminal actions and ensure public safety by identifying objects that have been left unattended in public spaces. For an object to be classified as abandoned, two conditions must be satisfied: the object must remain in one place and not be located near any individuals for a considerable length of time.

The input in abandoned object detection (AOD) systems includes images and videos captured from surveillance cameras in public spaces, such as airports, city squares and malls. This task can run either online, with real-time processing of video streams, or offline, where already captured video files are analysed at a later time. The OoI are typically bags, such as backpacks, handbags and suitcases. In addition to the detection of abandoned objects, the identification of their possible owner is also important. In this case, the system looks for the person that has been captured to hold the object for a period before abandoning it.

Intuitively, AOD systems assemble a number of sub-tasks that need to be completed in order to detect abandoned objects, including at first the identification of the OoI, then the tracking of their movement within the camera's range, and finally the decision about whether they have been abandoned. Thereby, the most effective AOD systems adopt a pipeline of modules that handle the aforementioned sub-tasks separately. For this scope, two fundamental modules in an AOD system are an object detector and an object tracking algorithm.

Various studies in the field of AOD have been conducted extensively, with many attempts to solve the problem using deep learning methods. Dwivedi et al. propose a solution to detect abandoned objects by first extracting foreground objects via background subtraction, and then categorising them as static or moving by comparing their position in consecutive frames. For their experiments, they mainly used the ABODA dataset.[31]

In Shyam et al., the authors suggest a pixel-based system to detect abandoned objects, which can perform in real-time.[32] They utilise sViBe for foreground object extraction, and subsequently a pixel-based finite state machine for stationary object detection. Lastly, they use a Single Shot MultiBox Detector (SSD) to recognise the OoI. The experiments were conducted using four public datasets: PETS2006,[33] PETS2007,[34] AVSS2007,[35] and ABODA.

A deep learning framework for the detection of abandoned bags is proposed in Sidyakin et al., consisting of a two-block process. In the first block, the Gaussian mixture model is used for pixel-level background modelling to output precise positions for the bounding boxes of the OoI. In the second block, a CNN architecture is used for bag recognition on a regional level.[36] Similarly, a two-stage method for luggage detection is proposed in Wojke et al. In this case, a motion estimation model is added in the first stage along with background subtraction, and a cascade CNN is utilised for the recognition of abandoned luggage, instead of a baseline CNN, as proposed in Smeureanu et al.[37]

In the context of the INFINITY project, a non-learning-based algorithm has been employed for abandoned object detection. This algorithm determines whether an object has been abandoned by analysing specific criteria. Firstly, it examines the state of the object, determining whether it is in motion or stationary. Subsequently, it evaluates whether the object is attended to by a person or if it is unattended. The use of a non-learning-based algorithm, instead of learning-based ones, provides a lightweight approach that can deliver both speed and accuracy in the results.

Figure 1: PETS2006 dataset representative samples

3. Future Challenges and Opportunities

Despite the significant progress that has been made in the field of visual content understanding in recent years, there are still many challenges to be addressed. Algorithms' scalability, interpretability, as well as data biases, are all examples of these challenges. In terms of scalability, a computer vision algorithm that has been trained in a specific dataset needs to perform well on a new set of unseen data. This is important for real-world applications, where the algorithm needs to be able to handle a variety of inputs and produce accurate results. As for interpretability, related issues occur when highly complex deep learning algorithms are employed for a task, making it difficult to understand how and why these algorithms make decisions.

Interpretability is important, especially for debugging and improving the algorithm, as well as building trust with end-users. Lastly, data bias issues occur when Machine Learning (ML) algorithms tend to favour certain types of data over others, leading to misleading and inaccurate results. The concept of bias in ML algorithms refers to the phenomenon where the model assigns importance to specific features in order to achieve better generalisation across larger datasets with diverse attributes, reducing sensitivity to individual data points. The datasets used for the training of these algorithms play a critical role as their volume, diversity, and relevance can impact the algorithm's performance. In other words, biases may arise if the datasets are skewed or not fully representative of the groups they are intended to serve. For example, in object detection and recognition, these issues can manifest as bias towards the identification of certain types of objects.

Additionally, while deep learning models have been shown to be highly effective at many visual content understanding tasks, they can be computationally expensive and require large amounts of training data. Therefore, addressing these challenges will require ongoing research and collaboration between experts in computer vision, machine learning, and related fields in order to ensure fair and accurate results.

Some technical challenges are presented below, pertaining to the problems associated with many visual content understanding tools:

- In object detection and tracking tasks, objects' size plays an important role in their detection and tracking, as small objects can be difficult to detect and track.
- In object tracking tasks, highly similar targets present a challenge for the relevant algorithms, making it difficult to maintain distinct object identities.
- In object tracking tasks, frequent occlusions can cause confusion in the detectors in crowded environments, as they make targets indistinguishable, resulting in gaps in their trajectories.
- In object detection tasks, weather conditions, varying illuminations or shadows cause inherent ambiguities when calculating the results.

Despite the above challenges, the existing technologies in the field of object detection and tracking (as well as visual content understanding in general) provide many opportunities for LEAs. One of the benefits is the ability to leverage efficient data-driven models and advanced algorithms to help identify patterns and trends in criminal activities faster, enabling LEAs to respond in a timely manner before threats arise. In other words, these technologies can improve the speed and accuracy of the investigations by providing access to large volumes of data and contributing to quickly narrowing down the search for evidence. Overall, these systems, powered by AI technologies, can be used to enhance surveillance capabilities, allowing LEAs to monitor activities in real-time, and subsequently, help them ensure public safety.

References

1. This Policy brief was prepared by CERTH, as part of T10.5.
2. Cf. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020), Yolov4: Optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934.
3. Cf. Guo, T., Dong, J., Li, H., & Gao, Y. (2017), "Simple convolutional neural network on image classification", in 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA) (pp. 721-724). IEEE.
4. Cf. Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022), "Masked-attention mask transformer for universal image segmentation", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1290-1299).
5. Cf. Gupta, N., Gupta, S. K., Pathak, R. K., Jain, V., Rashidi, P., & Suri, J. S. (2022), "Human activity recognition in artificial intelligence framework: A narrative review", in Artificial intelligence review, 55(6), 4755-4808.
6. Cf. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, and Zitnick, C. L. (2014), "Microsoft coco: Common objects in context", in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 740-755). Springer International Publishing.
7. Cf. Ding, Y., Hua, L., & Li, S. (2022), "Research on computer vision enhancement in intelligent robot based on machine learning and deep learning", in Neural Computing and Applications, pp. 1-13.
8. Cf. Wen, L. H., & Jo, K. H. (2022), Deep learning-based perception systems for autonomous driving: A comprehensive survey, in Neurocomputing.
9. Cf. Suganyadevi, S., Seethalakshmi, V., & Balasamy, K. (2022), "A review on deep learning in medical image analysis", in International Journal of Multimedia Information Retrieval, 11(1), pp. 19-38.
10. Cf. Khan, S. W., Hafeez, Q., Khalid, M. I., Alroobaea, R., Hussain, S., Iqbal, and Ullah, S. S. (2022), Anomaly detection in traffic surveillance videos using deep learning, in Sensors, 22(17), 6563.
11. Cf. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020), Yolov4: Optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934;
He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017), "Mask r-cnn", in Proceedings of the IEEE international conference on computer vision (pp. 2961-2969); and
Wojke, N., Bewley, A., & Paulus, D. (2017), "Simple online and realtime tracking with a deep association metric", in 2017 IEEE international conference on image processing (ICIP) (pp. 3645-3649). IEEE.
12. Cf. Mao, J., Shi, S., Wang, X., & Li, H. (2022), 3D object detection for autonomous driving: a review and new outlooks. arXiv preprint arXiv:2206.09474.
13. Cf. Lorenčík, D., & Zolotova, I. (2018), "Object recognition in traffic monitoring systems", in 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA) (pp. 277-282). IEEE.
14. Cf. Pienaar, S. W., & Malekian, R. (2019), "Human activity recognition using visual object detection", in 2019 IEEE 2nd Wireless Africa Conference (WAC) (pp. 1-5). IEEE.
15. Cf. 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).
16. Cf. Redmon, J., & Farhadi, A. (2017), "YOLO9000: better, faster, stronger", in Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).
17. Cf. Redmon, J., & Farhadi, A. (2018), Yolov3: An incremental improvement, ARXIV PREPRINT ARXIV:1804.02767.
18. Cf. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020), Yolov4: Optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934.
19. Cf. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, and Zitnick, C. L. (2014), "Microsoft coco: Common objects in context", in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 740-755). Springer International Publishing.
20. Cf. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014), „Rich feature hierarchies for accurate object detection and semantic segmentation", in Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).
21. Cf. Girshick, R. (2015), "Fast r-cnn", in Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

22. Cf. Ren, S., He, K., Girshick, R., & Sun, J. (2015), Faster r-cnn: Towards real-time object detection with region proposal networks, in Advances in neural information processing systems, 28.
23. Cf. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017), "Mask r-cnn", in Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).
24. Cf. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014), „Rich feature hierarchies for accurate object detection and semantic segmentation", in Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).
25. Cf. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014), „Rich feature hierarchies for accurate object detection and semantic segmentation", in Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587); and Girshick, R. (2015), "Fast r-cnn", in Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).
26. Cf. Ren, S., He, K., Girshick, R., & Sun, J. (2015), Faster r-cnn: Towards real-time object detection with region proposal networks, in Advances in neural information processing systems, 28.
27. Cf. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017), "Mask r-cnn", in Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).
28. Cf. Ren, S., He, K., Girshick, R., & Sun, J. (2015), Faster r-cnn: Towards real-time object detection with region proposal networks, in Advances in neural information processing systems, 28.
29. Cf. Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016), "Simple online and realtime tracking", in 2016 IEEE international conference on image processing (ICIP) (pp. 3464-3468). IEEE.
30. Cf. Wojke, N., Bewley, A., & Paulus, D. (2017), "Simple online and realtime tracking with a deep association metric", in 2017 IEEE international conference on image processing (ICIP) (pp. 3645-3649). IEEE.
31. Cf. Dwivedi, N., Singh, D. K., & Kushwaha, D. S. (2020), „An approach for unattended object detection through contour formation using background subtraction", in Procedia Computer Science, 171, 1979-1988; For the ABODA dataset see: https://github.com/kevinlin311tw/ABODA
32. Cf. Shyam, D., Kot, A., & Athalye, C. (2018), "Abandoned object detection using pixel-based finite state machine and single shot multibox detector" in 2018 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.
33. https://ftp.cs.reading.ac.uk//pub/PETS2006/
34. https://ftp.cs.reading.ac.uk//pub/PETS2007/
35. http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html
36. Cf. Sidyakin, S. V., & Vishnyakov, B. V. (2017), "Real-time detection of abandoned bags using CNN", in Automated Visual Inspection and Machine Vision II (Vol. 10334, pp. 149-160). SPIE.
37. Cf. Wojke, N., Bewley, A., & Paulus, D. (2017), "Simple online and realtime tracking with a deep association metric", in 2017 IEEE international conference on image processing (ICIP) (pp. 3645-3649). IEEE; and Smeureanu, S., & Ionescu, R. T. (2018, September), "Real-time deep learning method for abandoned luggage detection in video", in 2018 26th European Signal Processing Conference (EUSIPCO) (pp. 1775-1779). IEEE.