An Assistive Technology Based on Object Detection for Automated Task List Generation

Frédéric Rayar[a]
LIFAT, University of Tours, Tours, France

Keywords: Object Detection, YOLOv7 Model, Task Generation, Assistive Technology, Intellectual Disability.

Abstract: People suffering from Intellectual Disability (ID) face challenges to perform sequential tasks in their daily life, impacting their education and employability. To help them, the usage of Assistive Technology (AT) on mobile devices is a promising direction. However, most of them do not take advantage of the important recent advances in Artificial Intelligence. In this paper, we address this lack by presenting a prototype of an embed AT system that leverage computer vision advances to assist a user suffering from ID to perform sequential tasks in a guesthouse rooms' tiding-up activity. It first relies on a state-of-the-art object detector, namely YOLOv7, that we have adapted for a real-time usage on mobile devices. Then, using a "spot the difference" approach, it identifies objects of interest that are either present, absent or displaced, compared to a template image. A list of tasks to be achieved is automatically generated and conveyed to the user using an accessible and ergonomic interface. Early qualitative experiments of this ongoing work lead us to believe that our contribution could improve the life of people suffering from ID, allowing them to improve both their functioning and independence.

1 INTRODUCTION

Intellectual disability (ID) is a neurodevelopmental disorder, where one has limitations in intellectual functioning (such as learning, problem solving, judgement) and adaptive functioning (daily life activities such as communication and independent living).[1] These limitations could be either mild, moderate, severe or profound, affecting a person at different levels. ID is often associated with a genetic syndrome, for instance, the Down syndrome, the Fragile X syndrome or the Cornelia de Lange syndrome. Due to these limitations, performing sequential tasks alone is a rather challenging task for children, teenagers and adults suffering from ID. This has a direct impact on their education and employability, often leading them to social or economic exclusion.

Assistive technology (AT) concerns products (devices or software), whose primary purpose are to maintain or improve an individual's functioning and independence. It has a positive impact on the health and well-being of a person and his/her family, as well as broader socio-economic benefits, according to the World Health Organisation (WHO).[2] Its usage and benefits for people with ID is still currently being advocated (Boot et al., 2017) In recent years, mobile devices such as smartphones and tablets have been widely used as AT by providing relevant software solutions for people with a disability. Their usage in the context of ID is qualified as an *"emerging promise"* (Skogly Kversøy et al., 2020), while stating that *"it will open [their] world up"* (Danker et al., 2023). Leveraging recent advances in computer vision and artificial intelligence to develop mobile intelligent systems that can be used as AT appears to be promising. Indeed, object detection for extracting valuable information from images and videos can assist people with ID understand their surrounding environment and perform sequential tasks by themselves. Even if numerous object detection models have been proposed in the literature based on Convolutional Neural Networks (CNN), exploiting their outcomes for building AT systems remains a challenging task for the research community.

In this paper, we report our ongoing efforts to address this challenge. The main contributions can be summarised as follows: first, we adapted YOLOv7

[a] https://orcid.org/0000-0003-1927-8400

[1] https://www.psychiatry.org/patients-families/intellectual-disability/what-is-intellectual-disability

[2] https://www.who.int/news-room/fact-sheets/detail/assistive-technology

Rayar, F.
An Assistive Technology Based on Object Detection for Automated Task List Generation.
DOI: 10.5220/0012453200003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 4: VISAPP*, pages 601-605
ISBN: 978-989-758-679-8; ISSN: 2184-4321

(You Only Look Once) model (Wang et al., 2023), a widely used object detector to be used in real-time on mobile devices. Second, we proposed an algorithm to perform a *”spot the difference”* task between a query image and a template image. Third, our system can automatically generate a task list that can be provided to an end-user with ID via a simple and ergonomic interface. Finally, we have implemented a prototype for a specific usecase, namely the daily room tide up in guesthouses. The validation of the performed tasks, gamification for providing an engaging AT and user evaluation are not discussed here as they are still ongoing.

The paper is organised as follows: Section 2 presents a brief overview of the related works. The proposed approach is detailed in Section 3. Section 4 presents the adressed usecase and the current implementation of our AT system. Finally, we conclude and present directions for future works.

2 RELATED WORKS

2.1 Image Comparison

In terms of image comparison, a *”spot the difference”* task can be formulated either as an image subtraction or an image matching problem, which are also standard problems in computer vision. The purpose of image subtraction (Paranjape, 2000) is to enhance the differences between two images : after a step of image registration, the resulting image is obtained by subtracting one image from another, at pixel level. This simple and fundamental techniques has mostly been used in time-domain astronomy to study astronomical objects change in time. Close to this field is change detection (Singh, 1989), that refers to the process of identifying differences in the structure or properties of objects on Earth by analysing two or more images taken at different times. Recent neural-network have been considered to address change detection such as CNN (de Jong and Sergeevna Bosman, 2019) or Transformers (Bandara and Patel, 2022). One can find a thorough survey in (Parelius, 2023). However, these approaches mainly deal with remote sensing images and require a complex image registration step.

On the other hand, if we consider the image matching paradigm, the problem boils down to establishing a sufficient number of pixel or region correspondences from two or more images. Classic algorithms use the optical flow (Beauchemin and Barron, 1995), that computes a vector field that maps each pixel from one image to a corresponding pixel in another image, while others algorithms match salient image points (such as Harris corner (Harris and Stephens, 1988), SIFT (Lowe, 2004), SURF (Bay et al., 2006)) with pruning and robust fitting methods. Recent approaches have used deep learning, by considering the matching at CNN layers’ feature representation (Chen and Heipke, 2022). However, the complexity and underlying overhead of these approaches make them not fitted for real-time usage in mobiles devices.

2.2 Object Detection

Another approach to perform the *”spot the difference”* task is proceed at instance-level and therefore consider objects of interest in images . Object detection has always been an important field in the domain of computer vision. If one focus on the recent neural-network-based paradigm, this problem has been addressed by two different approaches:

- with a two-stage approach: first, identify regions of interest where objects are expected to be found, then detect objects in those regions and eventually refine their bounding boxes. Most famous algorithms that follows this two-stage approach are R-CNN (Girshick et al., 2014) and Fast(er) R-CNN (Ren et al., 2017);
- with a one-stage approach: here, a fully convolutional approach is used and the network is able to find all objects within an image in only one pass. Most popular algorithms with this approach are YOLO (You Only Look Once) (Redmon et al., 2016) and SSD (Single-Shot Multibox Detector) (Liu et al., 2016).

The two-stage approaches usually have a slightly better accuracy but are slower to run, while single-shot algorithms are more efficient and have as good accuracy, explaining why the latter options are mostly used nowadays.

3 PROPOSED APPROACH

Figure 1 illustrates the workflow that we propose in this paper. First an image is acquired via a mobile device such as a smartphone or a tablet. Second a real-time efficient detection algorithm is used to recognize a predefined list of objects in the captured image. Third, a *”spot the difference”* task is performed between the query image and template images that have been stored in a database beforehand. Then, a task list is automatically generated and delivered to

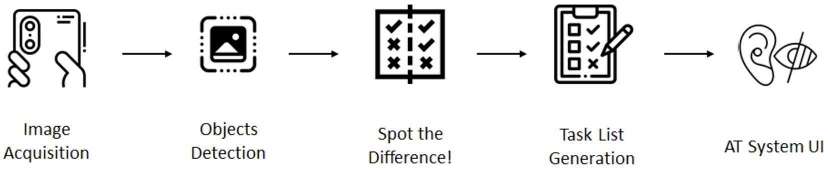


Figure 1: Proposed workflow to perform automatic generation of task list leveraging object detection[3].

the end-user using accessibility considerations. These steps are detailed below.

3.1 Embedded YOLO

In our study, we have chosen the YOLO model (Redmon et al., 2016) over SSD (Liu et al., 2016). This choice is based on detection accuracy and speed criteria, considering that we want to embed this object detector in a real-time AT mobile system. No preprocessing to deal with classic challenges in real-scene image acquisition, such as image quality restoration or image luminosity rectification, is performed on the current prototype. More specifically, we have chosen a YOLOv7 (Wang et al., 2023) open source C++ implementation and adapted it for an Android system[4].

YOLOv7 recognises all objects that are present in the MS-COCO dataset (Lin et al., 2014) and since most of them are not relevant in the considered use-case, it affects the smoothness and hence the usability of the AT system. Therefore, we build a functionality that allows an administrator of the system to select the list of objects that is relevant to a given scenario. The speed up ratio when considering only 6 objects of interest over 80 object categories is 6, and allows to reduce the visual overload that could have a severe impact on people suffering form ID.

3.2 Spot the Difference

We detail here the proposed algorithm to perform a *"spot the difference"* task. Assume a template image I, that contains a list of objects $O = \{O_1, O_2, \ldots, O_n\}$ and their corresponding bounding boxes $BB = \{BB_1, BB_2, \ldots, BB_n\}$. We recall that a bounding box is given by $BB = (x, y, w, h)$, with (x, y) the top-left coordinates of the box and (w, h) the width and the height of the box, respectively. For a given query image Q, we assume that the object detector algorithm has found the list of objects $FO = \{FO_1, FO_2, \ldots, FO_m\}$ and their bounding boxes $FBB = \{FBB_1, FBB_2, \ldots, FBB_m\}$.

```
Data: $O = \{O_1, O_2, \ldots, O_n\}$,
      $FO = \{FO_1, FO_2, \ldots, FO_m\}$
Result: $presence$
$presence.reset()$;
for $i \leftarrow 1$ to $n$ do
    $obj \leftarrow O_i$;
    for $j \leftarrow 1$ to $m$ do
        $f\_obj \leftarrow FO_j$;
        if $obj.label == f\_obj.label$ then
            $presence[i] \leftarrow True$;
        end
    end
end
```

Algorithm 1: Presence/absence detection of an object.

```
Data: $O = \{O_1, O_2, \ldots, O_n\}$,
      $BB = \{BB_1, BB_2, \ldots, BB_n\}$,
      $FO = \{FO_1, FO_2, \ldots, FO_m\}$,
      $FBB = \{FBB_1, FBB_2, \ldots, FBB_m\}$,
      $iou\_threshold$
Result: $displacement$
$displacement.reset()$;
for $i \leftarrow 1$ to $n$ do
    $obj \leftarrow O_i$;
    $bb \leftarrow BB_i$;
    for $j \leftarrow 1$ to $m$ do
        $f\_obj \leftarrow FO_j$;
        $f\_bb \leftarrow FBB_j$;
        if $obj.label == f\_obj.label$ then
            $iou \leftarrow compute\_iou(bb, f\_bb)$ ;
            if $iou < iou\_threshold$ then
                $displacement[i] \leftarrow True$;
            end
        end
    end
end
```

Algorithm 2: Displacement detection of an object.

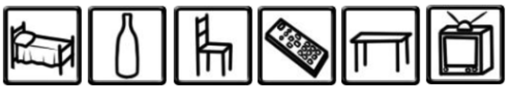
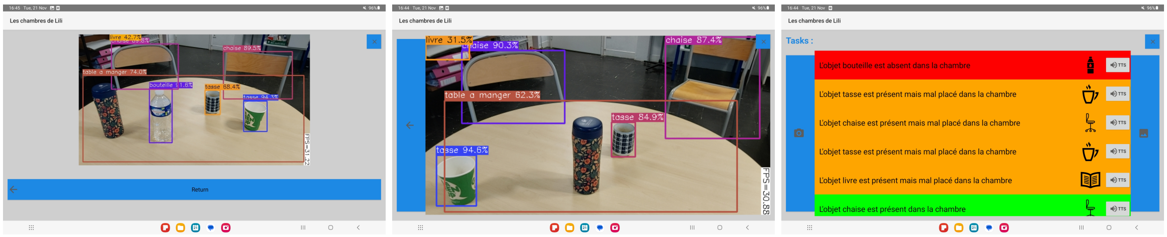


Figure 2: List of default objects to detect in our current implementation. Images are from the MS-COCO website[5].

[3] This illustration uses resources from Flaticon.com

[4] https://github.com/xiang-wuu/ncnn-android-yolov7

[5] https://cocodataset.org/#explore



L'objet bouteille est absent dans la chambre

L'objet tasse est présent mais mal placé dans la chambre

L'objet chaise est présent mais mal placé dans la chambre

L'objet tasse est présent mais mal placé dans la chambre

L'objet livre est présent mais mal placé dans la chambre

L'objet chaise est présent dans la chambre

Figure 3: Examples of three interfaces of the proposed AT system: (left) template image with detected objects, (middle) query image, after some objects have been removed or displaced and (right) automatically generated task list with ergonomic and accessible considerations.

Two levels of spotting have been considered: the first detects the absence or presence of a given object in Q, compared to the template image I, and the second, more refined, consider if an object in Q is displaced compared to the template image I. Algorithms 1 and 2 details the two proposed spotting approaches, respectively. For the displacement detection, we have considered the Intersection over Union (IoU) metric, that is used to evaluate the performance of object detectors. It computes the extent of overlap of a ground truth bounding box BB_i to a detected bounding box FBB_j, using the following formula:

$$\text{IoU} = \frac{\text{Area of Overlap}(BB_i, FBB_j)}{\text{Area of Union}(BB_i, FBB_j)}.$$

The greater the region of overlap, the greater the IoU value. In this work, we have empirically set a threshold value $iou\_threshold = 0.5$, to estimate the displacement of an object. Finally, by identifying the presence, absence or misplacement of objects of interest, a list of tasks to fit the template image I is naturally deduced.

4 PROPOSED AT SYSTEM

4.1 Usecase Description

In our first implementation, we have identified a specific usecase, namely the rooms' tiding-up activity in guesthouses. Each room has been segmented in specific sections (*e.g.* desk part, bed part, bathroom, etc.) and template images of these sections have been stored beforehand. To speed up the *"spot the game"* task, the templates images have been processed by the object detector algorithm and relevant information have been stored. The proposed AT system aims at supporting a person suffering from ID in the following way: *(i)* the user enters a room, *(ii)* he position himself at a given section of the room and takes a picture, *(iii)* the object detection is performed on a given subset of objects, *(iv)* knowing the room and the section, the *"spot the difference"* task is performed between the query image and template images, *(v)* the system then automatically generates the list of tasks to be performed in order to tidy up the room. More specifically, this list is created according to the presence or displacement of objects of interest, in order to restore the room in its initial stage, as depicted in the template image, and welcome new guests. Finally the AT system (in)validates tasks that have been performed by asking the user to take another picture (on-going works, not describe in this paper). Figure 2 illustrates the objects that are relevant in our usecase and that are recognized by default by the current implementation of our AT system (bed, bottle, chair, remote, table, tv).

4.2 Prototype

Figure 3 illustrates examples of three interfaces of the proposed AT system, in a test environment. The left picture shows a template image, where the following objects were found: a bottle, two cups, 2 chairs, a table and a book. Note that the book is only detected with only a 42.7% confidence and that the designed thermo-insulated bottle is not detected. The middle picture presents a query image, where some objects were removed or displaced: one bottle was removed, one cup and one chair were displaced. Finally the right picture presents the view of our generated task list.

In order to provide an accessible and ergonomic system, we have considered the following considerations: *(i)* the possibility to customise the font, *(ii)* a simple yet consistent palette (red, orange and green) to represent the object status (absent, displaced and present, respectively), *(iii)* the usage of pictograms, commonly used in Augmentative and Alternative Communication (AAC) and *(iv)* the usage of an audio feedback, on-demand. The validation of the tasks to be performed can be done in the current version of the proposed AT system, by retaking successive pictures, however we believe that the gamification of this validation process could increase the user engagement, justifying our ongoing work.

5 CONCLUSION AND PERSPECTIVES

In this paper, we have presented ongoing work on the development of an AT system that leverage recent advances in AI, to assist people suffering from ID to perform sequential tasks autonomously in a guesthouse rooms' tiding-up activity. It relies on the cutting-edge YOLOv7 object detector, that is used to identify objects of interest that need to be moved. A task list is automatically generated and conveyed to the user using an accessible and ergonomic interface. In future research, two main directions have been identified and are ongoing : *(i)* the validation of the performed tasks, through the gamification of the proposed AT system, to ensure the engagement if the user and *(ii)* the on-site user evaluation of the prototype, involving thus young adults in a real usage for the room's tidying-up in guesthouse.

REFERENCES

Bandara, W. G. C. and Patel, V. M. (2022). A transformer-based siamese network for change detection. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210.

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. In *Computer Vision – ECCV 2006*, pages 404–417.

Beauchemin, S. S. and Barron, J. L. (1995). The computation of optical flow. *ACM Comput. Surv.*, 27(3):433–466.

Boot, F. H., Dinsmore, J., Khasnabis, C., and MacLachlan, M. (2017). Intellectual disability and assistive technology: Opening the gate wider. *Frontiers in Public Health*, 5:5–10.

Chen, L. and Heipke, C. (2022). Deep learning feature representation for image matching under large viewpoint and viewing direction change. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:94–112.

Danker, J., Strnadová, I., Tso, M., Loblinzk, J., Cumming, T. M., and Martin, A. J. (2023). 'it will open your world up': The role of mobile technology in promoting social inclusion among adults with intellectual disabilities. *British Journal of Learning Disabilities*, 51(2):135–147.

de Jong, K. L. and Sergeevna Bosman, A. (2019). Unsupervised change detection in satellite images using convolutional neural networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587.

Harris, C. G. and Stephens, M. J. (1988). A combined corner and edge detector. In *Alvey Vision Conference*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, pages 740–755.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 21–37.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110.

Paranjape, R. B. (2000). 1 - fundamental enhancement techniques. In BANKMAN, I. N., editor, *Handbook of Medical Imaging*, Biomedical Engineering, pages 3–18. Academic Press.

Parelius, E. J. (2023). A review of deep-learning methods for change detection in multispectral remote sensing images. *Remote Sensing*, 15(8).

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(06):1137–1149.

Singh, A. (1989). Review article digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10(6):989–1003.

Skogly Kversøy, K., Kellems, R. O., Kuyini Alhassan, A.-R., Bussey, H. C., and Daae Kversøy, S. (2020). The emerging promise of touchscreen devices for individuals with intellectual disabilities. *Multimodal Technologies and Interaction*, 4(4).

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.