An object detection approach for blind and visually impaired people.

Authors : Raj Bhensadadia* (201501011), Bansi Gajera* (201501028),
Dhwani Mehta* (201501054), Kanan Vyas* (201501121)

Email Id : raj.b.btechi15@ahduni.edu.in, bansi.g.btechi15@ahduni.edu.in,
dhwani.m.btechi15@ahduni.edu.in, kanan.v.btechi15@ahduni.edu.in

*School of Engineering and Applied Science, Ahmedabad University

***Abstract*—Computer vision technologies have been rapidly developed in recent years. It is promising to use the state-of-art computer vision techniques to help people with vision loss.The system is able to handle occlusions, sudden camera/object movements, rotation or various complex changes. Finally, an object classification module is proposed that exploits the YOLO algorithm and extends it with new categories specific to assistive devices applications. The experimental evaluation, performed on the Coco dataset and on a set of videos acquired by us, demonstrates the effectiveness and efficiency of the proposed method.**

***Keywords*--Navigation, computer vision, distance estimation, object detection, object recognition**

I. INTRODUCTION

The World Health Organization estimates there are about 314 million vision impaired people in the world, of which about 45 million are blind. The leading causes of blindness are cataract, uncorrected refractive errors, glaucoma, and macular degeneration. Many people who are seriously vision impaired use a white cane and/or a guide dog to avoid obstacles. Moving through an unknown environment becomes a real challenge when we can't rely on our own eyes. Since dynamic obstacles usually produce noise while moving, blind people develop their sense of hearing to localize them.

Recent statistics, relative to people with visual disabilities published by the World Health Organization (WHO) in August 2014, show that more than 0.5% of the total population suffers from visual impairments (VI). Among these, 39 million people are completely blind. Unfortunately, by the year 2020 worldwide the number of individuals with VI is estimated to double.

In Section 2, we briefly review the state of the art. Section 3 presents the proposed cognition system that involves two major stages: obstacle detection and tracking. The experimental results, conducted on the Coco dataset. Finally, Section 5 concludes the paper and opens new directions for further work.

II. RELATED WORK

Most often, they concern trained dogs or white canes. Today, the white cane always represents the simplest and most affordable travel aid available. Within this context, the elaboration of an assistive device dedicated to blind and visually impaired people that can improve cognition over the environment and facilitate the safe, autonomous navigation in novel outdoor spaces is a crucial challenge. In this paper, we propose an assistive device that combines computer vision techniques and deep convolutional neural networks in order to detect, track and recognize objects encountered during the outdoor navigation.

The focus is put on assistive systems, based on computer vision methods, dedicated to the visually impaired users. The experimental results, conducted on the Coco dataset as well as on a video corpus acquired in real life scenarios are presented.

The system works in real-time and proves to be robust to ego-motion and noise.

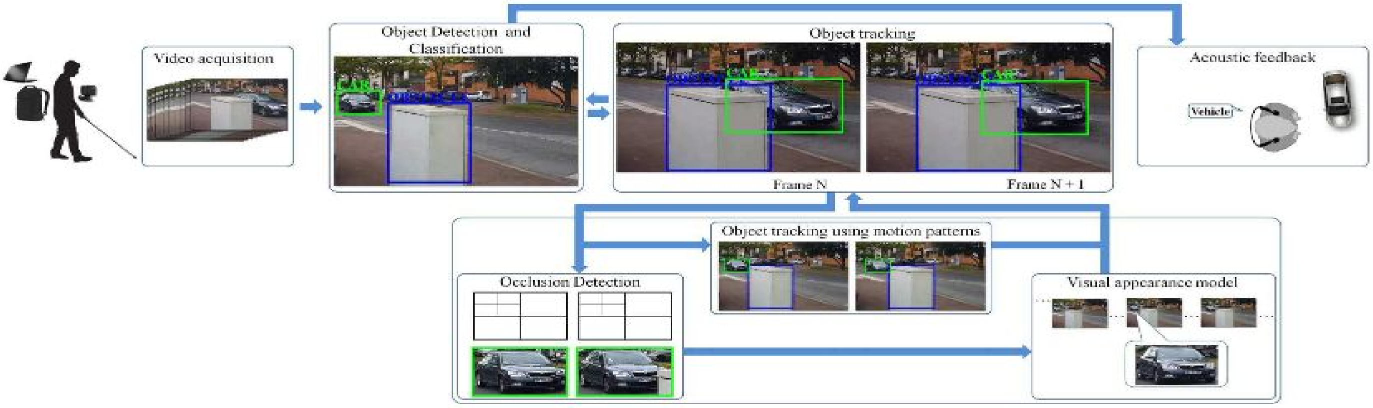


Figure 1: The global architecture of the proposed approach.

III. PROPOSED APPROACH

1) Object Detection

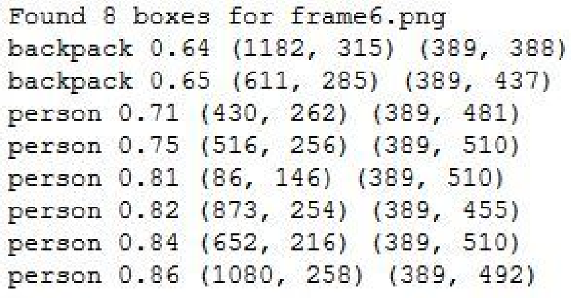
In recent years, the tracking-by-detection approaches have become increasingly popular for solving the problem of robust object localization in subsequent video frames, despite important object motion, changes in view-point or other acquisition-related variations.

The initial object detection is performed by applying the YOLO algorithm on the first frame of the video stream. YOLO treats the object detection problem as a regression mechanism for spatially separated bounding boxes and their associated class probabilities. The authors decided to use YOLO due to the real-time processing capabilities and its reduced number of false positives.

We also implemented object detection algorithm using faster r-cnn and single shot detector. The results of faster r-cnn was more accurate then YOLO but in comparision of speed YOLO was much faster then faster-rcnn.

In addition, the detector can be used to predict candidate location for novel objects in video frames where such action is required.

In addition, the detector can be used to predict candidate location for novel objects in video frames where such action is required.



```
Found 8 boxes for frame6.png
backpack 0.64 (1182, 315) (389, 388)
backpack 0.65 (611, 285) (389, 437)
person 0.71 (430, 262) (389, 481)
person 0.75 (516, 256) (389, 510)
person 0.81 (86, 146) (389, 510)
person 0.82 (873, 254) (389, 455)
person 0.84 (652, 216) (389, 510)
person 0.86 (1080, 258) (389, 492)
```

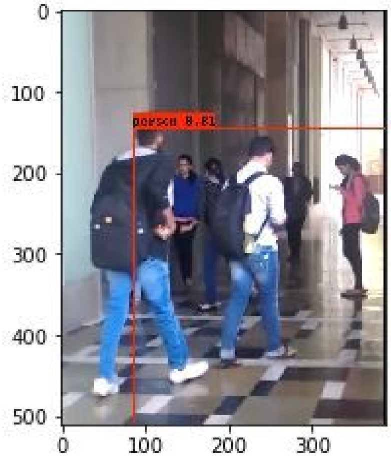


Fig. 2. Object detection using YOLO

2) Object Tracking

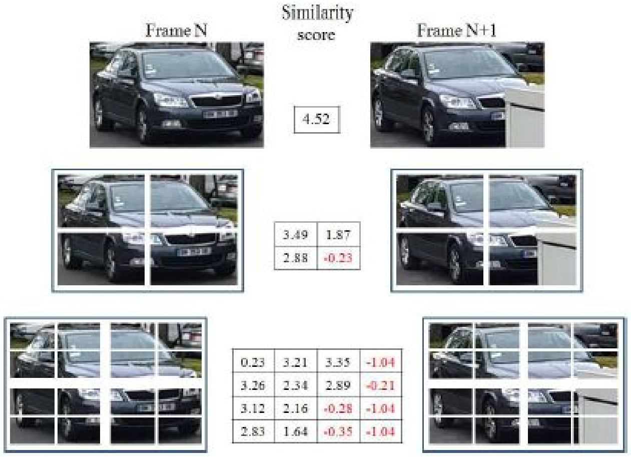
The proposed approach is a generic object tracker based on two convolutional neural networks trained offline. In this way, no online fine tuning is performed.The network receives as input the target object as well as the associated search regions.

The output is a set of high level image features that are applied as input to the fully connected layers. The tracking system based on motion patterns proves to be very fast , robust and accurate . In addition,the object estimated position together with its associated context area is insufficient to reliably determine if the new location of the bounding box actually contains the object of interest. In order to overcome such limitations, Authors have proposed to integrate in the process rich visual cues, established from the object previous positions and appearances.

After obtaining the initial candidate location, Authors have introduced a refinement strategy that aims to adaptively modify the bounding box position and shape in order to avoid incorrect/false object tracks due to the background clutter. The refinement process includes two stages, which are occlusion detection and object appearance modeling. The process of occlusion detection and handling is illustrated in Fig. Authors have applied a quadtree decomposition algorithm in order to divide the candidate object location and its reference bounding box into a set of non-overlapping image patches.
The partition process is repeated until the third level of decomposition. We decided to use only three levels of decomposition in order to ensure a reasonable degree of descriptiveness of the similarity measure. The image patches, at the initial resolution and from all levels of decomposition, are compared against the correspondent one in the reference frame. The similarity degree between the image patches is obtained using the DeepCompare algorithm.
The comparison technique is a CNN-based model trained to take into account a wide variety of changes into the image appearance. The system does not require any manually tuned features and is able to learn, directly from the training data, a general similarity function that serves to compare patches. The image patches are processed by using a 2-channel network architecture that offers the best trade-off between the computational speed and the system accuracy. The two patches being compared are considered as a 2-channel image that is directly applied to the first convolutional layer of the neural network.
The bottom of the CNN is composed of a series of convolutional, ReLU and max-pooling layers. The top module is a fully connected, linear decision layer. The system has great flexibility and is fast to train. As in , we propose to divide the convolutional layers into smaller 3x3 kernels separated by ReLU activations.

$$MS_{score} = \max\left\{S_{score_{cut_1}}; S_{score_{cut_2}}\right\};$$

Considering the object as being in an occluded state if the associated for the image patches situated on the second and third level of decomposition return negative values Also, objects of interest characterized by larger bounding boxes show a similar behavior . The process consists in reducing the size of the bounding box with 1/8 of the initial size.
The selection of 1/8 of the initial size makes it possible to avoid too brutal shrinkage of the bounding box.



4.52

3.49	1.87
2.88	-0.23

0.23	3.21	3.35	-1.04
3.26	2.34	2.89	-0.21
3.12	2.16	-0.28	-1.04
2.83	1.64	-0.35	-1.04

Fig.3. Occlusion detection using quad-tree decomposition

In addition, no cut is allowed for image patches with less than 5 pixels on the third level of decomposition. The proposed tracker is considerably more effective than a regular tracker based solely on a strong motion model. Various trackers, based on visual features construct appearance models for both, the interest objects and the background information. Due to the real-time constraint imposed on our application, we decided to develop a model solely for the tracked objects.
Commonly, most trackers use a single/fixed appearance model selected from the first frame of the video stream. A

single model is insufficient to cope with important changes in obstacles shape, pose or features. In order to overcome this limitation, a continuous update of the object appearance model is required. In the state of the art, various authors consider as a positive example the tracker's current location and attempt to predict the object novel position within a neighborhood search area, by exploiting the object's trajectory information .

Even though this approach shows promising results, it suffers from several drawbacks. Thus, if the tracker is not sufficiently precise when estimating the novel object location, the object appearance model tends to be updated with sub-optimal positive examples . In contrast, if multiple positive examples are selected from nearby locations, the object model is constantly updated and the current appearance can incorporate too much contextual information and thus become confusing . In our work, we have adopted a tracking-by-detection approach that continuously updates the object appearance model with novel instances whenever such an action is required.

The objective is to estimate, with high accuracy, the new position of the object bounding box in the adjacent frame. The input is the candidate location returned by the motion-based tracking algorithm . The predicted object position, together with its associated context region is further analyzed for a more accurate object location estimation. The context region is subsequently used as a search area in order to determine, independently, the best location for each instance in the object appearance model.

At each stage, the similarity score provided by the DeepCompare algorithm is computed. In order to reduce the processing burden instead of a brute force search, we have adopted a hierarchical approach, similar to the block-based motion estimation in method used in MPEG-4 . The location that yields the highest Deep Compare similarity score is retained as correct for the current object appearance model. To validate the object location, we impose the maximum similarity score to be superior to the average score obtained within the temporal sliding window that incorporates the last video frames processed.

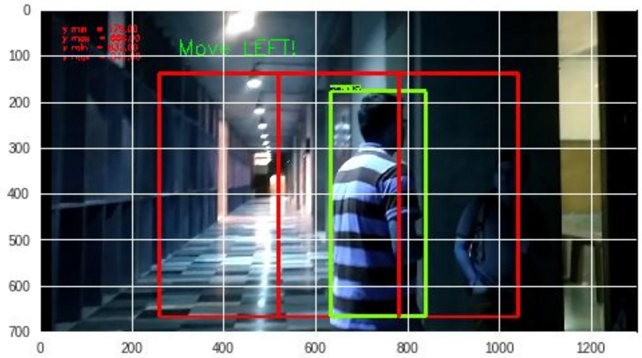
The object appearance model is constantly updated with novel elements if the visual similarity scores of the current instance with all the frames being analyzed satisfy the similarity condition. In the same time, the most ancient object instance in the model is discarded. In this way, we ensure that the object appearance is not updated with sub-optimal instances .

3) Obstacle Classification

The image patches are classified using a modified version of the YOLO algorithm. We extended the system with additional training classes. The object class is predicted by performing a global reasoning about the entire video frame.

4)Acoustic feedback module

We observed that not all obstacles presented in the scene represent a potential risk for the blind. A detected object is marked as urgent if it is situated within the proximity of interest , otherwise the obstacles is categorized as normal or non-urgent. By employing two areas of proximity we can prevent the system to launch acoustic warning messages for all the detected objects existent in the scene.



IV. EXPERIMENTAL RESULTS

Datasets and baseline systems

All the sequences were annotated by human observers. The videos are acquired at a resolution of 320 x 240 pixels, are trembled and cluttered.

Implementation details

All the experiments were performed on a Google Colab.

Evaluation measures

In addition, the measure completely ignores the interest object size and does not take into account the apparent tracking failure. We evaluated the proposed tracker using the quantitative measures, as described below. This measure is computed as the overlap between the predicted target region and the ground truth annotation data .

$$\Phi = \{\phi_t\}_{t=1}^{N}, \quad \phi_t = \frac{R_t^G \cap R_t^T}{R_t^G \cup R_t^T};$$

$$\phi_t = \frac{R_t^G \cap R_t^T}{R_t^G \cup R_t^T} = \frac{TP}{TP + FN + FP};$$

V. CONCLUSION

In this paper, Authors have proposed a novel perception system based on computer vision methods and deep convolutional neural networks able to assist the visual impaired user during the outdoor navigation.

In contrast to various techniques existent in the state of the art our system is able to detect, track and recognize, in real-time, all relevant object existent in the scene without any apriori knowledge about shape, position or dynamics. The output of the system is transformed into a set of acoustic warnings transmitted to the Visually impaired users.

VI. REFERENCES

[1] World Health Organization (WHO) - Visual impairment and blindness. Available online: http://www.who.int/mediacentre/factsheets/fs282/en/

[2] A. Rodríguez, J.J. Yebes, P.F. Alcantarilla, L. M Bergasa; J. Almazán, "*Assisting the Visually Impaired: Obstacle Detection and Warning System by Acoustic Feedback Sensors 2012"*, 12, 17476-17496.

[3] R. Tapu, B. Mocanu and E. Tapu, "*A survey on wearable devices used to assist the visually impaired user navigation in outdoor environments,*" 2014 11th International Symposium on Electronics and Telecommunications (ISETC), Timisoara, 2014, pp. 1-4.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. "*You only look once: Unified, real-time object detection"*. In
CVPR, 2015.

[5] M. Kristan, R. Pflugfelder et al. *The visual object tracking VOT2016 challenge results*. In ECCV Workshop, pp. 1–45, 2016.

[6] Leo, M.; Medioni, G.G.; Trivedi, M.M.; Kanade, T.; Farinella, G.M. *Computer Vision for Assistive Technologies*. CVIU. 2017 , 154, 1–15

[7] S. Cloix, V. Weiss, G. Bologna, T. Pun and D. Hasler, "*Obstacle and planar object detection using sparse 3D information for a smart walker*," International Conference on Computer Vision Theory and Applications (VISAPP) 2014, pp. 292-298.

[8] R. Manduchi, "*Mobile vision as assistive technology for the blind: An experimental study".* In. ICCHP' 2012.

[9] R. Tapu, B. Mocanu, A. Bursuc and T. Zaharia, "*A Smartphone-Based Obstacle Detection and Classification System for Assisting Visually Impaired People*," In ICCV-Workshops, 2013, pp. 444-451.

[10] B. Mocanu, R. Tapu, T. Zaharia, *"When Ultrasonic Sensors and Computer Vision Join Forces for Efficient Obstacle Detection and Recognition".* Sensors 2016, 16, 1807.

[11] J. Manuel Saez, F. Escolano and A. Penalver, "*First Steps towards Stereo-based 6DOF SLAM for the Visually Impaired,*" In CVPR - Workshops, 2005, pp. 23-23.

[12] V. Pradeep, G. Medioni and J. Weiland, "*Robot vision for the visually impaired*," In CVPR - Workshops, 2010, pp. 15-22.