Chapter 10
Digital Enhancement of Cultural Experience and Accessibility for the Visually Impaired

Dimitris K. Iakovidis, Dimitrios Diamantis, George Dimas, Charis Ntakolia, and Evaggelos Spyrou

10.1 Introduction

Today, approximately 16% of the world's population lives with some form of visual impairment (WHO, 2018). Individuals with low or total absence of vision have to deal with various daily problems, struggling to fit in the modern way and rhythm of life. To address this important issue, researchers in the fields of medicine, smart electronics, computer science and engineering are joining their forces to develop assistive systems for the visually impaired individuals. To date, as a result of this effort, several designs and components of wearable camera-enabled systems for the visually impaired have been proposed.

A survey of relevant systems proposed until 2008 has been presented in Zhang, Ong, and Nee (2008). It identifies three categories of navigation systems: (a) based on positioning systems, including Global Positioning System (GPS) for outdoor positioning, and pre-installed pilots and beacons emitting signals, e.g., radiofrequency, infrared (IR), ultrasonic, etc., to determine the absolute position of the user in a local structured environment, (b) based on Radiofrequency Identification (RFID) tags with contextual information, such as surrounding landmarks, turning points; and vision-based systems exploiting the information acquired from digital cameras. In a more recent study performed in the beginning of 2017 (Elmannai

D. K. Iakovidis (✉) · D. Diamantis · G. Dimas · C. Ntakolia
Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece
e-mail: diakovidis@uth.gr; didiamantis@uth.gr; gdimas@uth.gr; cntakolia@uth.gr

E. Spyrou
Institute of Informatics and Telecommunications, National Center for Scientific Research "DEMOKRITOS", Athens, Greece
e-mail: espyrou@iit.demokritos.gr

S. Paiva (ed.), *Technological Trends in Improved Mobility of the Visually Impaired*,
EAI/Springer Innovations in Communication and Computing,
https://doi.org/10.1007/978-3-030-16450-8_10

& Elleithy, 2017), the state-of-the-art sensor-based assistive technologies were reviewed and assessed. The conclusions of that work indicate that most of the current solutions are still at a research stage, partially solving the problem of either indoor or outdoor navigation. It also suggests some guidelines for the development of relevant systems, which include: (a) real-time performance, i.e., fast processing for the exchanged information between the user and the sensors, and detection of suddenly appearing objects within a range of 0.5–5 m, regardless of the place and time; (b) wireless connectivity, (c) reliability, (d) simplicity, (e) wearability, and (f) low cost, affordable for most users.

Focusing on the most recent vision-based systems, their main, most critical functionalities include the detection of obstacles and provision of navigational assistance, whereas additional features include the recognition of objects, or scenes in general. A wearable mobility aid solution based on embedded 3D vision was proposed in Poggi and Mattoccia (2016). By wearing this device the users can perceive, be guided by audio messages and tactile feedback, receive information about the surrounding environment and avoid obstacles along a path. Another relevant system was proposed in Schwarze et al. (2016). That system was capable of perceiving the environment with a stereo camera, providing information about the obstacles and other objects to the user in the form of intuitive acoustic feedback (through sonification of objects/obstacles adjacent to the user). A system for joint detection, tracking and recognition of objects encountered during navigation in outdoor environments was presented in Tapu, Mocanu, and Zaharia (2017). The key principle considered for the development of that system was the alternation between tracking using motion information and prediction of the location of an object in time based on visual similarity. A project exploiting a smart-glass was presented in Suresh, Arora, Laha, Gaba, and Bhambri (2017). It investigated the development of a system that consists of a camera and ultrasonic sensors to recognize obstacles ahead, and assess their distance in real-time. The processing was performed on a portable computer. A wearable camera system proposed in Wang, Katzschmann, et al. (2017) was capable of providing also haptic-feedback to the user through vibrations. It was capable of identifying walkable spaces, planning a safe motion trajectory in the space, as well as recognition and localization of certain types of objects. A system called Sound of Vision was presented in Caraiman et al. (2017), aiming to provide the users with a 3D representation of the environment around them, conveyed by means of the hearing and tactile senses. The vision system was based on an RGB Depth (RGB-D) sensor, with an Inertial Measurement Unit (IMU) was used for tracking the head/camera orientation. In Lin, Lee, and Chiang (2017) a simple smartphone-based guiding system was proposed. That system included a fast feature recognition module running on the smartphone for fast processing of visual data. It also included remotely accessible modules, one for more demanding feature recognition tasks, and one for direction and distance estimation.

An augmented reality system, featuring obstacle localization was proposed in Yu, Yang, Jones, and Saniie (2018). That system was using predefined augment reality markers to identify specific accessible facilities, such as hallways, restrooms, staircases and offices within indoor environments. A scene perception system was proposed in Kaur and Bhattacharya (2018), based on a multi-modal fusion-based framework for object detection and classification. In Yang, Wang, et al. (2018) a unifying terrain awareness framework was proposed as an extension of a basic vision system based on an IR RGB Depth (RGB-D) sensor (Yang et al., 2017), aiming at attaining efficient semantic understanding of the environment. The approach was integrated into a wearable navigation system by incorporating a depth segmentation method. Another vision-based navigational aid based on an RGB-D sensor was presented in Lin, Wang, Yang, and Cheng (2018); however, that study was focusing on a specific component for road barrier recognition.

A relevant pre-commercial system promising both obstacle detection and audio-based user communication is investigated in the context of an H2020 funding scheme for Small Medium Enterprises (SMEs). The system, called EyeSynth (Audio-Visual System for the Blind Allowing Visually Impaired to See Through Hearing)[1], is based on a stereoscopic imaging system mounted on a pair of eyeglasses and the audio signals communicated to the user are non-verbal and abstract. The implementation details are not yet available. Other relevant commercially available solutions include ORCAM MyEye,[2] which is attachable to the users' eyeglasses and discreetly reads printed and digital text aloud from printed or digital surfaces, and recognizes faces, products, and money notes; eSight Eyewear,[3] which aims to enhance the vision of partially blind individuals by using a high-speed, high-definition camera that captures whatever the user is looking at, and then displays it on two near-to-eye displays; AIRA system,[4] which connects blind or low-vision people with trained, remotely located human agents who have access to what the user sees through a wearable camera, at the touch of a button, e.g., in the case of an emergency. These commercially available solutions do not yet incorporate any intelligent components for automated assistance.

The review performed reveals that during the last 2 years several studies and research projects have been initiated, setting higher standards toward a system for computer-assisted navigation of the visually impaired individuals. This chapter presents the concept of a novel vision-based system being developed in the recently initiated project ENORASI (Intelligent Audiovisual System Enhancing Cultural Experience and Accessibility, 2018–2021, funded by European Union and Greek national funds). It describes state-of-the-art (and beyond) methods considered for its development, and it investigates the user requirements based on the relevant literature.

The rest of this chapter consists of six sections. Section 10.2 presents the concept of the proposed system. Section 10.3 focuses on methods investigated for the implementation of a *Computer Vision* (*CV*) system capable of artificially perceiving the users' environment. Section 10.4 describes the concepts related to the methods

[1] https://eyesynth.com.

[2] https://www.orcam.com.

[3] https://www.esighteyewear.com.

[4] https://aira.io/.

considered for the implementation of its interactive intelligent user interface and decision-making modules. In Sect. 10.5, a set of user requirements are mined from the relevant literature. Section 10.6 discusses the technologies that better adapt to the goals of the proposed system, and the last section summarizes the conclusions of this study.

10.2 Vision-Based Navigation in Outdoor Cultural Environments

Museum (indoor) accessibility for the visually impaired individuals has been investigated in several studies (Alkhafaji, Fallahkhair, Cocea, & Crellin, 2016; Shah & Ghazali, 2018). However, the accessibility of outdoor sites of cultural interest has attracted less attention, although the experiences from visiting such sites can be equally significant. The ENORASI project aims to investigate and deliver a pre-commercial digital system to assist the visually impaired individuals on moving safely in external environments of cultural interest, e.g., of historic value, while providing them an enhanced touring experience. Besides the audible guidance and instructions for obstacle avoidance, the system provides also information about the sights in a descriptive way, through an emotionally aware, intelligent, speech user interface.

The main components of the proposed system include (Fig. 10.1): (a) stereoscopic CV system for depth assessment, through visual sensors embedded on the users' eyeglasses; (b) emotion-aware speech interaction through a microphone and earphones that are also embedded to the users' eyeglasses in a way that it does not interfere with their hearing; (c) communication with a GPS-enabled Mobile Processing Unit (MPU), such as a smartphone or a tablet, customized for visually impaired individuals. A challenge is to enable robust performance in an energy-efficient way, based solely on visual sensors, without augmentation from additional sensors of the local environment such as ultrasound and IMU sensors. Also, to further increase autonomy, software optimizations, such as smart management of energy and computational resources are considered (Gubbi, Buyya, Marusic, & Palaniswami, 2013).

The proposed system is based on image/video and audio/speech processing and analysis methods. These include computer vision algorithms for automatic object recognition, e.g., obstacles, and the estimation of their distance from the user, and emotionally-aware speech recognition algorithms, which as well as algorithms for decision making based on the acquired multimodal data (images, audio, GPS). The analysis of the user experiences in relation to their emotions at different locations can be useful as a resource for feedback from the users to the system administrators, so as to enhance their services at these locations.

The processing and analysis of the acquired data is performed partially in the MPU, while more complex computational processes are performed in a remote

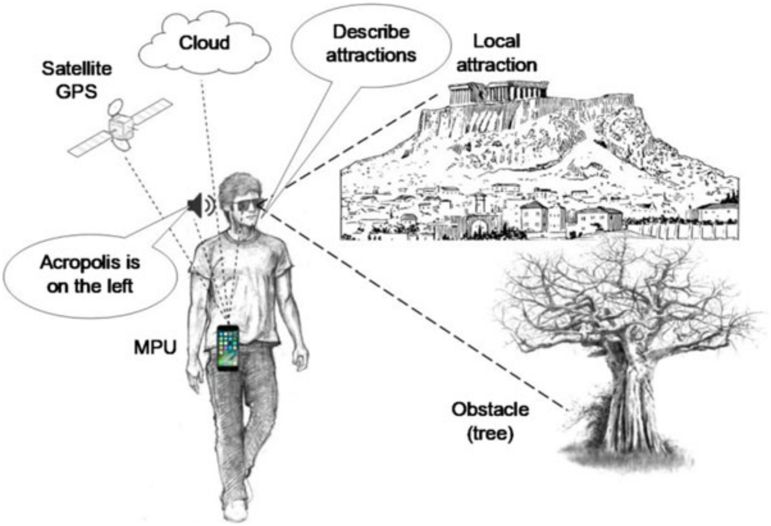

Fig. 10.1 The proposed system concept

server, through a computational cloud environment. Tasks performed in the MPU include obstacle detection, as it is the most critical task for the users' safety, as well as critical speech-based communication with the user, so as to enable basic functionalities even if the system is offline (i.e., the cloud is not accessible). More computationally demanding tasks, such as object or scene recognition, and decision making with respect to planning of the navigation route, or other higher-level inference such as route planning, complex speech and emotion recognition are performed on remote servers accessible through the cloud.

10.3 Human Vision Via Computer Vision

CV is a field of computer science that combines image processing and artificial intelligence to enable computers to recognize and assess the semantics within the images and video sequences. With its theoretical foundations back in the early sixties (Roberts, 1963), its developments have provided a variety of useful applications, spanning from everyday apps, such as the face detection feature of conventional cameras (Wang, Hu, & Deng, 2018), to specialized applications with social impact, such as image-guided anomaly detection in the medical domain (Iakovidis, Georgakopoulos, Vasilakakis, Koulaouzidis, & Plagianakos, 2018). The rapid evolution of parallel hardware architectures, such as the *Graphics Processing Units* (*GPUs*), has triggered unprecedented developments in *Artificial Neural Networks* (*ANNs*). These inherently parallel computational structures can bring us closer than ever to the development of systems that perceive the visual world like humans, and interpret it into auditory information for the visually impaired.

In this context, aspects investigated with respect to the perception of the visual world include the detection of obstacles, the recognition of objects, as well as the estimation of object sizes and distances.

10.3.1 Artificial Neural Networks for Computer Vision

Supervised ANN architectures, such as the *Multi-Layer Perceptron* (*MLP*) (Theodoridis & Koutroumbas, 2009), have been widely applied in the field of CV for object detection and recognition. In this context, usually shallow architectures, composed of three layers have been considered, with reference to their universal approximation capabilities (Hornik, Stinchcombe, & White, 1989). The input of these architectures was usually composed of so-called 'hand-crafted' image features, extracted using predefined methodologies. However, this imposed limitations in the generality of the CV approaches developed. To cope with such limitations, a revolutionary extension of the MLP for image classification, named *Convolutional Neural Network* (*CNN*), was presented in 1995 (LeCun, Bottou, Bengio, & Haffner, 1998). The core components of a CNN network are its convolutional layers, which contain a bio-inspired neuron connection arrangement mimicking the biological cells of visual cortex. In this layer, each neuron has pre-fixed connections to the input space, which form the so-called *receptive field* of the neuron. Multiple neurons span across the image with fixed receptive field and shared weights, which result into the extraction of the same feature across the entire input space, forming a *feature map*. Multiple feature maps are used to extract different features from the input space. This feature extraction process is applied over several convolutional layers. Also, pooling operations, such as maximum and average pooling, are performed after one or more convolutional layers, which facilitate dimensionality reduction. After this process, the resulting feature representations are classified by an MLP composed of three fully connected neuronal layers. The original CNN architecture proposed in LeCun et al. (1998) is known as LeNet. Due to the relatively large number of layers composing a CNN, its architecture is characterized as *deep*. This, along with the fact that such architectures are trainable, motivated the term *deep learning*, which is widely used to characterize machine learning using *Deep Neural Networks* (*DNNs*).

The CNN concept was revived in 2012, with an architecture named AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), extending the original CNN approach (LeCun et al., 1998). This network won the ImageNet ILSVRC-2012 competition, which involved classification experiments on a large dataset, composed of one million images with over 10,000 object categories. That network had 62.3 million parameters, and its training became feasible by exploiting GPU computing. Since then, CNNs have been widely adopted in various applications, including object recognition (Szegedy, Ioffe, Vanhoucke, & Alemi, 2017), detection (Lin, Goyal, Girshick, He, & Dollár, 2018), segmentation (He, Gkioxari, Dollár, & Girshick, 2017) and tracking (Held, Thrun, & Savarese, 2016). Considering their high computational requirements, a research branch is being developed toward the reduction of their complexity (Howard et al., 2017; Zhang, Zhou, Lin, & Sun, 2017), aiming to enable their introduction into mobile and embedded devices.

Unsupervised CNN architectures have been investigated as feature extractors in the form of *AutoEncoders* (*AE*) (Luo, Li, Yang, Xu, & Zhang, 2018). In principle, an AE consists of an input layer, called *encoder*, hidden layers, and an output layer, called the *decoder*. By training the network in an unsupervised manner, the resulting representation of the input space into its hidden layers form the image features, subsequently used for image classification. Another CNN type, called *Generative Adversarial Network* (*GAN*) (Goodfellow et al., 2014), is capable of generating image data by a reverse approach to that of AEs; instead of compressing a high-dimensional input space into its layers, it receives a low-dimensional vector which is subsequently used to generate a realistic output image. Useful applications in the context of CV include image resolution enhancement (Ledig et al., 2017) and visual saliency prediction (Pan et al., 2017), where the GAN is used to generate a saliency map, i.e., a map of regions within the images where objects of interest might be located.

Other ANN architectures that have been proved useful in the context of CV include the *Recurrent Neural Networks* (*RNNs*) and their extension, called *Long Short-Term Memory* (*LSTM*) (Hochreiter & Schmidhuber, 1997). RNNs maintain their internal hidden states to model the dynamic temporal behavior of sequences with arbitrary lengths through directed cyclic connections between its units. LSTMs extend RNNs by adding three gates to RNN neurons; namely, a so-called *forget gate* to control whether to forget the current state, an *input gate* to indicate if it should read the input, and an *output gate* to control whether to output the state. Recent approaches include combinations of these networks with CNNs for multi-label image classification (Wang et al., 2016), and video action recognition (Wang, Gao, Song, & Shen, 2017).

The following paragraphs provide further information on the state-of-the-art methods, including ANN architectures, considered in the context of this study.

10.3.2 Obstacle Detection

Obstacle detection addresses the detection of any object interfering with the motion trajectory of an agent, including a robot, a smart-vehicle, or a visually impaired person following the directions provided by a smart navigation system. In the following, an overview of the state-of-the-art object detection methods applicable in the context of obstacle detection for the visually impaired is provided.

Object Detection Methods

An integrated CNN-based framework for object detection was presented in Sermanet et al. (2013). That framework combined a CNN architecture for feature extraction based on AlexNet (Krizhevsky et al., 2012), named OverFeat, and a regression network to detect multiple bounding boxes around objects in images. A *Region-based* CNN architecture for object detection was presented in Girshick, Donahue, Darrell, and Malik (2014) with the name R-CNN. The methodology uses selective search (Uijlings, Van De Sande, Gevers, & Smeulders, 2013) to extract 2000 class-agnostic region proposals from each image, which are then resized and feed-forwarded into a pre-trained CNN model to extract features. The extracted features are then used to train a linear *Support Vector Machine* (*SVM*) classifier (Theodoridis & Koutroumbas, 2009) which classifies the extracted feature representations. Although R-CNN outperformed the OverFeat approach (Sermanet et al., 2013) for object detection, it requires more computational resources. To reduce its computational complexity, the Fast R-CNN (Girshick, 2015) was proposed, in which feature maps are extracted from the entire input image. From these feature maps, region proposals are extracted and reshaped into a fixed size, by a technique called Region of Interest (RoI) pooling, so that they can be processed by a fully connected layer. The Softmax function is used to predict the class of the RoI vector while in parallel it computes the offset values for the bounding box of the object.

Another architecture, called *Spatial Pyramid Pooling Network* (*SPPNet*) (He, Zhang, Ren, & Sun, 2015) aimed to cope with the problem of the fixed-size input required by the CNNs which may impact the detection accuracy of the overall model. This was done by implementing a novel spatial pyramid pooling which enabled the network to generate fixed-length image representation regardless of the image size. Compared with R-CNN, SPPNet relies on the same principles, yet it does not have to process 2000 region proposals per image, as R-CNN does. Each bounding box is classified by an SVM and bounding box regressor. A Faster R-CNN (Ren, He, Girshick, & Sun, 2015) achieved real-time object detection capabilities, by removing the selective search used by the previous methodologies.

A methodology for object detection that is fundamentally different from the previous ones was presented in Redmon, Divvala, Girshick, and Farhadi (2016). It is called *You Only Look Once* (*YOLO*) and it relies solely on a single forward pass of an input image. The image is subdivided using a fixed-size grid, and entered to a CNN that predicts bounding boxes and class probabilities for each box. A saliency-inspired neural network model for object detection was proposed in Erhan, Szegedy, Toshev, and Anguelov (2014). It predicts a set of class-agnostic bounding boxes along with a single score for each box, corresponding to its likelihood of containing any object of interest. In Liu, Anguelov, et al. (2016) an object detector with name *Single Shot multibox Detector* (*SSD*) which achieved good balance between computational performance and prediction accuracy. A region-based, *Fully Convolutional Network* (*FCN*: a CNN without fully connected layers) was proposed in Dai, Li, He, and Sun (2016). It relies on the generation of position-sensitive score maps to cope with the dilemma between translation-invariance in image classification and translation-variance in object detection. In Lin et al. (2017) an object detector for multi-scale object detection was proposed. That detector relies on a feature extractor, named *Feature Pyramid Network* (*FPN*), which was designed to improve detection accuracy and speed. In Redmon and Farhadi (2017) YOLO9000, an extension of the YOLO approach (Redmon et al., 2016), was introduced for real-time object detection, considering 9000 object categories. A single-shot object detector, named *Deconvolutional SSD* (*DSSD*), was presented in Fu, Liu, Ranga, Tyagi, and Berg (2017). It extended SSD by replacing the original VGGNet with a *Residual Network* (*ResNet*) (He, Zhang, Ren, & Sun, 2016) for feature extraction. ResNet architecture relies on small building blocks, named residual blocks, that feature skip connections and simple Convolutional-ReLu-Convolutional layers, which result in a network with 152 layers.

RetinaNet, proposed in Lin, Goyal, et al. (2018), is a single, unified network composed of a backbone network and two task-specific sub-networks. The backbone network is implemented by a ResNet architecture, used for feature extraction. The first sub-network performs the classification and the second one performs bounding box regression. A multi-scale extension of the DSSD network, called *Multi-Scale Deconvolutional SSD* (*MDSSD*), has been proposed in Cui (2018), specifically for small object detection.

Obstacle Detection for the Visually Impaired

The ability to detect different types of objects in images is crucial for a system aiming to assist the navigation of the visually impaired. In the context of the project ENORASI, the user has to be able to trust the detection system to detect multiple types of obstacles/objects of different sizes in real-time, while in parallel the system should be able to accurately and reliably detect the surrounding area for potential cultural sights. Although some of the reviewed deep learning approaches are able to tackle the issue of real-time object detection, they are computationally demanding. This increases the need for a robust, multi-scale object detector, able to perform real-time object detection in mobile devices, so that the users will not have to rely on a client-server detection model that can degrade the overall performance due to network latency.

Among the various object detection methods that have been applied in the context of CV-based obstacle detection for the visually impaired, this paragraph focuses on the most recent ones. The obstacle detection module of the wearable mobility aid proposed in Poggi and Mattoccia (2016) was based on LeNet. The object detection in the DEEP-SEE framework presented in Tapu et al. (2017) was based on a YOLO CNN (Redmon et al., 2016). For the smart-glass approach presented in Suresh et al. (2017) three CNN architectures were encountered, namely Faster R-CNNs (Ren et al., 2015), YOLO (Redmon et al., 2016) and SSDs (Liu, Anguelov, et al., 2016). A Faster R-CNN (Ren et al., 2015) was used to detect and track objects in Kaur and Bhattacharya (2018). Motion, sharpening and blurring filters were used to enhance feature representation.

A state-of-the-art multi-scale FCN that we proposed in Diamantis, Iakovidis, and Koulaouzidis (2019) is presented as a candidate to cope with efficient obstacle detection. The proposed architecture, called Look-Behind FCN (LB-FCN), features multi-scale feature extraction capabilities with Look-Behind (LB) residual connections (Fig. 10.2a). The multi-scale feature extraction is bundled in a single block named Multi-scale Convolutional Block (MCB). Several MCBs are connected together forming a deep FCN (Fig. 10.2b). The LB connections aim to preserve the input volume, along with the extracted features per MCB. Each MCB input volume is feed-forwarded through the LB connection to the output of the MCB module where it is aggregated using an addition operation. The resulting network combines advantages of state-of-the-art architectures, including ResNet (He et al., 2016), ResNeXt (Xie, Girshick, Dollár, Tu, & He, 2017) and Inception-v4 (Szegedy et al., 2017) (discussed in the next section) but it features a lower number of free parameters, contributing to its time-efficiency over these and conventional CNNs,

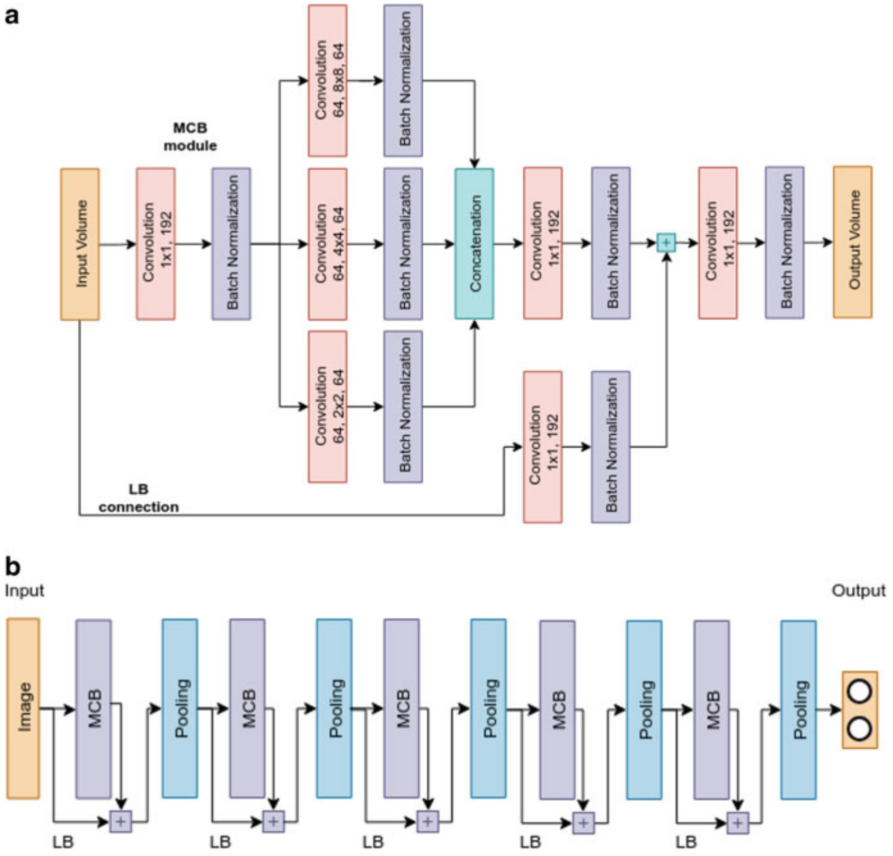

Fig. 10.2 The LB-FCN architecture. (**a**) The basic component of the architecture, formed by the MCB module and the LB connection. (**b**) An example LB-FCN with five MCB-LB components

such as VGGNet (Simonyan & Zisserman, 2014). Also, due to its multi-scale feature extraction capability it enables object detection at various scales. LB-FCN was benchmarked on open-access medical datasets, outperforming the state-of-the-art architectures and methods. Also, experiments using different datasets for training and testing of the architecture indicate its robustness against diversity of the objects to be detected and its generalization potentials (Diamantis, Iakovidis, & Koulaouzidis, 2018).

10.3.3 Object and Scene Recognition

Besides the detection of obstacles, which is critical for the safety of the visually impaired, object or scene recognition provides an additional quality in the visual perception of the world that can influence the decisions of the subjects during a guided tour. An object recognition system can provide information about the type of an obstacle, e.g., distinguish if the obstacle is a human or a tree, about the presence of a cultural sight within a scene and identify it, e.g., identify the Parthenon or the Caryatids of the Erechthion monuments in Acropolis. Object or scene recognition can be considered as a computationally more complex extension of object detection, since an intelligent system, such as an ANN, has to incorporate additional free parameters to encode additional knowledge about the different object types to be recognized. This means that most of the ANN-based object detection approaches reviewed in the previous subsection are extensible for object recognition. Similarly, ANN architectures proposed for object recognition can be simplified for object detection. In the following an overview of the most recent approaches to object recognition, applicable in the context of the proposed system, is provided.

Generic Recognition Approaches

Most of the state-of-the-art object recognition systems are also based on CNN architectures. Today, the Visual Geometry Group Network (VGGNet), and its variation VGG-16 (Simonyan & Zisserman, 2014), is considered as a baseline approach. VGGNet-16 is a CNN composed of 16 trainable layers with 138 million free-parameters. GoogLeNet CNN architecture (Szegedy et al., 2015), also known as Inception-v1, has a design that makes use of multiple Inception modules. An Inception module consists of parallel convolutional layers, with multi-scale feature extraction capabilities. Another CNN architecture that has been used for object recognition is ResNet (He et al., 2016). This architecture, which was also studied for object detection, won the ILSVRC-2015 challenge and for the first time, surpassing the human classification top-5 error rate by 5–10%.

A revised version of GoogLeNet architecture (Inception-v1) was presented in Szegedy, Vanhoucke, Ioffe, Shlens, and Wojna (2016) with name "Inception-v2", aiming to reduce the computational complexity by lowering the number of free parameters of the network. To achieve that, the network utilized factorized convolutions along with aggressive regularization. With primary focus on increasing the computational efficiency of DNN architectures, a recent study (Iandola et al., 2016) presented SqueezeNet architecture. The network was able to provide AlexNet-level accuracy on ImageNet dataset but with 50 times less number of free parameters. To formalize the Inception series architectures and to investigate if residual learning can benefit Inception-like architectures, a series of networks was presented in Szegedy et al. (2017). The result was the creation of three networks named Inception-v4, Inception-ResNet-v1 and Inception-ResNet-v2, with the first being a pure Inception architecture while the following Inception and ResNet hybrids. ResNeXt was presented in Xie et al. (2017) as an enhanced sequel of ResNet, expanding the original residual module with multiple parallel convolutional layers. The number of parallel convolutional layers in each ResNeXt building block, characterized as the *cardinality* of the network and after series of experiments it proved to be an equally important hyper-parameter when designing a network. ResNet was also the source of inspiration for the DenseNet (Huang, Liu, Van Der Maaten, & Weinberger, 2017) architecture. This architecture is based on a series of Dense Blocks which contain a series of convolutional layers, each one connected with all the following layers of the module in a feed-forward fashion. To battle the problem of long training time required by ResNet architecture Huang, Sun, Liu, Sedra, and Weinberger (2016) presented a Deep Network with Stochastic Depth. The network was trained utilizing a novel methodology similar to the dropout layer. Upon training, instead of disabling a percentage of neurons in a single layer, stochastic depth training disables entire layers of the network, vastly decreasing the training time. As an added benefit the authors found that training with stochastic depth can positively affect the classification accuracy of the network. To reduce the number of free parameters and allow CNN architectures to be used on mobile and embedded devices, Howard et al. (2017) presented a series of networks named *MobileNets*. The architecture utilizes depth-wise separable convolutions, instead of conventional convolutional layers which reduce the computational complexity. MobileNets expose two hyper parameters, named width and depth multipliers that balance the trade-off between the accuracy and the computational efficiency. Aiming to the same goal as MobileNets, Zhang, Zhou, et al. (2017) presented a CNN architecture with name *ShuffleNet*. The authors followed point-wise group convolution and channel shuffling to reduce the computational cost and maintain high classification accuracy. Experiments presented show ShuffleNet can achieve AlexNet classification accuracy on ImageNet dataset, increasing the speed by 13 times.

The semantic interpretation of scenes can also be considered as part of a scene recognition system. CNN architectures proposed for this purpose include encoder–decoder architectures, such as *SegNet* (Badrinarayanan, Kendall, & Cipolla, 2017).

Object or Scene Recognition for the Visually Impaired

Those of the recent vision-based navigation systems (Sect. 10.1) featuring object or scene recognition are based on CNN architectures as well. The mobility aid solution proposed in Poggi and Mattoccia (2016) uses a LeNet architecture for categorization of objects in eight classes. A kinetic real-time convolutional neural network for navigational assistance was presented by Lin, Wang, et al. (2018) with name *KrNet*. The system relies on a CNN architecture designed to provide navigational assistance for visually impaired individuals in the problem of road barrier recognition. The terrain awareness framework proposed in Yang, Wang, et al. (2018) was based on a CNN for semantic image segmentation. Various CNNs were tested including SegNet (Badrinarayanan et al., 2017).

Beyond the state-of-the-art approaches, the computational complexity of the multi-scale LB-FCN architecture that we proposed in Diamantis et al. (2019) (Sect. 10.3.2.2, can be further reduced by applying the depth-wise separable convolution approach proposed in Howard et al. (2017), and extended for multi-label classification of objects as described in Vasilakakis, Diamantis, Spyrou, Koulaouzidis, and Iakovidis (2018), so as to enable the recognition of multiple objects. Considering the benchmarks performed in Diamantis et al. (2019), it constitutes a promising alternative for the time-efficient object detection and recognition in the time-critical context of the system presented in this chapter.

10.3.4 Visual Distance Estimation

The research field tackling with the problem of the estimation of the traveled distance of a subject, based exclusively on visual cues, is known as Visual Odometry (VO). VO has been thoroughly investigated and approached by researchers from different perspectives (Forster, Zhang, Gassner, Werlberger, & Scaramuzza, 2017; Konda & Memisevic, 2015; Zhang, Kaess, & Singh, 2017) and on different application domains (Dimas, Spyrou, Iakovidis, & Koulaouzidis, 2017; Fang & Scherer, 2015; Maimone, Cheng, & Matthies, 2007). VO can be used to supplement or even replace other traditional navigation options, since it cannot be affected by GPS dropouts due to obstacles or other unfavorable conditions (Nistér, Naroditsky, & Bergen, 2004). VO methodologies can be rendered as an alternative navigational assistance method for the visually impaired individuals, since it can produce high-quality results with regard to the traveled distance approximation.

Quite a few works have been proposed, incorporating VO methods for the navigational assistance of the visually impaired. A framework involving multiple sensors for assistive navigation of the visually impaired was proposed in Xiao et al. (2015). That system features real-time localization by exploiting VO for the estimation of the location of the user with an RGB Depth (RGB-D) camera. The system proposed in Schwarze et al. (2016) was able to perceive the environment through a stereoscopic camera, using head tracking with visual odometry, an IMU sensor and sonification of objects/obstacles adjacent to the user. In another study (Aladren, López-Nicolás, Puig, & Guerrero, 2016) among multiple sensors tested, an RGB-D sensor was selected as sufficient. With the use of the RGB-D sensor both the depth and the visual information were sufficient for the detection of the main structural elements of a scene, in order to determine an obstacle-free path for the safe passage of a visually impaired individual.

The system proposed in Wang, Katzschmann, et al. (2017) segments the free space and maps it into free space motion instructions. In another study (Lin, Cheng, Wang, & Yang, 2018), robust visual localization (VO) is achieved via a GoogLeNet (Szegedy et al., 2016) and global optimization. To tackle the problem of accurate VO in crowded environments, an egocentric VO approach was proposed for crowd-resilient indoor assisted navigation (Yang, Duarte, & Ganz, 2018). A monocular VO approach for the assisted navigation of visually impaired individuals was proposed in Ramesh, Nagananda, Ramasangu, and Deshpande (2018). The aim of that work was to tackle the problem of real-time VO in indoor environments with a single camera. To achieve that, imaging geometry, VO, and object detection along with distance-depth estimation algorithms were combined.

Despite the interesting results reported by the afore-mentioned studies and the practical potentials of the deployment of the VO methodologies, there are still several challenges that need to be tackled. For example, the utilization of multiple sensors for better accuracy in navigation and distance estimation systems requires handling the synchronization among the sensors (Xiao et al., 2015). The positioning of the camera sensor is also important, since it can lead to unwanted noise in the collected data (Xiao et al., 2015). Another challenge to be tackled is the adaptability of the system in different environments, such as crowded, indoor and outdoor environments (Lin, Cheng, et al., 2018; Yang, Duarte, & Ganz, 2018). A direction toward coping with this issue is the use of machine learning algorithms, e.g., recently, *Recurrent CNNs* (*RCNNs*) have provided very good results in performing VO (Li, Wang, Long, & Gu, 2018; Wang, Clark, Wen, & Trigoni, 2017). CNNs have the capacity of learning optimal features for the task that they are trained to perform. Thus, a single RCNN may have the learning capability to extract features resilient to crowded, indoor and outdoor environments. Also, its recurrent nature enables making correlations between previous and next situations. However, using deep learning algorithms such as RCNNs, a lot of computational resources are needed, whereas real-time performance becomes also a challenge. Thus, further investigation and development of methodologies should include ways of handling computational payload on MPUs.

The proposed system exploits stereoscopic imaging for robust depth estimation, which can also contribute in more accurate VO, as compared with monocular VO. For this purpose the use of the state-of-the-art Intel® RealSense™ D435[5] sensor is investigated. This sensor is small enough (90 × 25 × 25 mm) to be mounted on the front side of the users' eyeglasses. It enables 3D depth sensing, with a maximum

[5] https://realsense.intel.com/intel-realsense-downloads/.

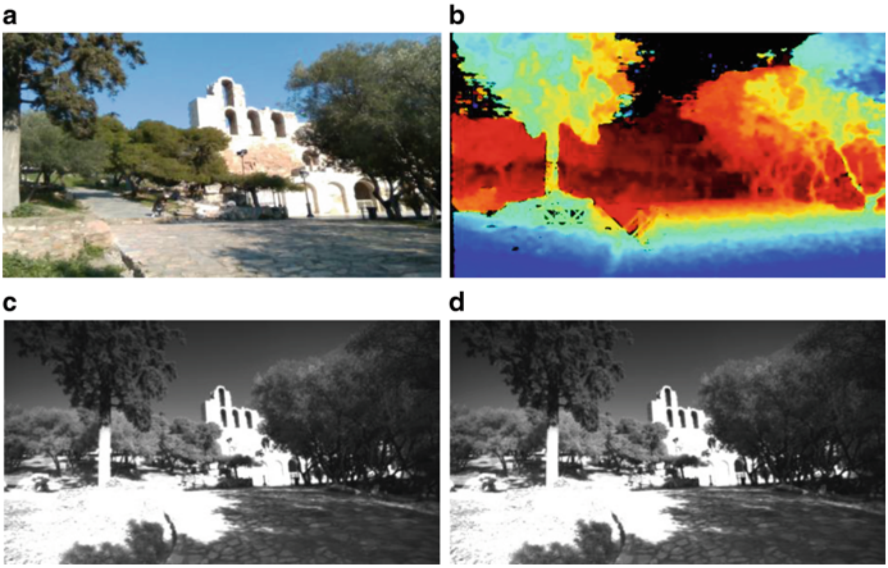

Fig. 10.3 Scene captured with the visual sensors of the proposed system near the Theatre of Herodes Atticus. (**a**) RGB image. (**b**) Depth map (more distant objects appear more reddish). (**c**) Left-stereo IR image. (**d**) Right-stereo IR image. The different visual sensors have different field of views

range of 10 m. It includes a stereoscopic system composed of two IR cameras, an IR projector and a high-resolution RGB camera. The infrared projector improves the depth estimation of the stereo camera system by projecting a static IR pattern on the scene. The IR pattern projection enables the texture enrichment of low texture scenes. An example of a scene captured with this sensor is illustrated in Fig. 10.3.

Visual Size Estimation

In the context of the proposed system, size estimation provides an added value in the assessment of obstacles and objects, which contributes in the enhancement of the users' experience. For example, it can be used as a cue to derive the level of deviation from a path, due to the presence of an obstacle, or as a quality to understand the size of a monument or a statue in a cultural environment.

The previous works in size estimation are limited. An object size measurement method that utilizes a stereo camera setup for the purposes of object identification has been proposed in Mustafah, Noor, Hasbi, and Azma (2012). That methodology includes object detection in the stereoscopic images, blob extraction and distance estimation. Another method aiming to both object localization and size measurement has been proposed in Liu, Yu, Chen, and Tang (2016). The algorithm used was based on the *Circle of Apollonius* for estimations without computing trigonometric functions. A review study on CV methodologies based on either monocular or stereo camera systems for size measurements was performed in Hao, Yu, and Li (2015). That review showed that the accuracy of the CV approach can be higher than traditional measurement methods. Another study (Pu, Tian, Wu, & Yan, 2016) proposed a novel framework to measure multiple objects in a scene using one or two photo shots.

More recently, we proposed a size measurement methodology that uses motion estimation over a video frame sequence in order to avoid the use of external references in size estimation (Iakovidis, Dimas, et al., 2018). This methodology exploits the distance estimated by an ANN toward the target object to be measured and then, the geometric model of the camera is used to estimate its dimensions.

There are still a lot of challenges to overcome toward accurate size measurements in the wild, using exclusively computer vision methodologies. As the authors of Hao et al. (2015) indicate in their review, the size estimation is prone to errors due to curved shape of the targeted object and to low/distorted image quality. In our work (Iakovidis, Dimas, et al., 2018), the size estimation is based on an accurate distance estimation from the camera to the target object, so the error is analogous to the distance estimation error. Also, the accurate object detection is of major importance for the good performance of size estimation methods (Mustafah et al., 2012).

In the proposed system, the visual size estimation via the method proposed by Iakovidis et al. (2018) can be augmented by the depth estimation provided by the use of the RGB-D sensor described in the previous subsection, as well as by the object detection methods discussed in Sect. 10.3.2. With the depth information and the accurate detection of the objects of interest, the geometric size measurement is possible after object segmentation.

10.4 Emotion-Aware Speech User Interface and Decision Making

In most mobile application human–machine interaction is performed through a visual user interface provided. In the case of visually impaired users, alternative, mainly auditory options, such as tactile and voice-enabled user interfaces, are preferable (Csapó, Wersényi, Nagy, & Stockman, 2015).

10.4.1 Voice User Interface

A Voice User Interface (VUI) aims to enable human–machine interaction through speech. Typically, VUIs are used to allow voice input as a means of controlling several types of devices or software applications. During the last few years, the advances of cloud services have led to the integration of VUIs in many popular mobile applications and smart environments. Smart assistants such as Apple's Siri, Amazon's Alexa, Samsung's Bixby, or Google Assistant have emerged. These are software agents integrated in smart mobile phones, smart-watches, or smart speakers. Their role includes performing tasks or services, e.g., managing emails/calendars using verbal commands, answering questions of users, e.g., regarding the news, or the weather and also controlling home automation, e.g., lights and thermostats. VUIs have also been playing a key role in several applications in call centers, where typically are used as interactive voice response systems, with limited natural language processing capabilities. Undoubtedly, VUIs have become very popular through their integration into devices of daily use, such as the smartphones.

10.4.2 Emotion Recognition

One of the most recent trends in the field of human–computer interaction is the recognition of the user's emotional state (Cowie et al., 2001). During the last few years, several approaches have been proposed, that are based on sensors, placed either within the users' environment such are cameras and microphones, or sensors that are wearable or embedded into devices carried by users, such as physiological or inertial sensors. In case of an approach based on the user's visual appearance, emotion recognition is based on facial, posture or motion features (Baltrusaitis et al., 2011; Piana, Stagliano, Odone, Verri, & Camurri, 2014). Body sensors may measure either physiological parameters, such as body temperature, heart and respiratory rate, muscle activity, skin conductance response or even brain activity (Haag, Goronzy, Schaich, & Williams, 2004) or extract inertial features (Tsatsou et al., 2018). Audio-based approaches are typically divided into two categories: (a) those that are based on the extraction of low- or mid-level features (Papakostas et al., 2017); and (b) those that are based on the processing of the spoken content, e.g., using a natural language understanding approach. It should be noted that each of these approaches has its own limitations, while user acceptance is typically low, e.g., in the case of cameras, the users may feel that their privacy is violated. Moreover, body sensors may cause discomfort when used for a long time. Totally non-invasive approaches are not yet available.

However, approaches that make use only of microphones are considered to be the least invasive. Sensors may be easily placed within the users' environment. Also, embedded microphones of mobile phones may be used. When assured that spoken content is not analyzed, and instead audio features are only used, people are less sensitive with privacy issues. Therefore, it should be clarified that the two discrete parts from which spoken content is composed. The *linguistic* content of speech includes the articulated patterns, as they are pronounced by the speaker. The *non-linguistic* content of speech may be described as the variation of the pronunciation of the aforementioned patterns, i.e., how the linguistic content has been pronounced (Anagnostopoulos, Iliou, & Giannoukos, 2015). When the goal is to classify spoken content to its underlying emotions based on its non-linguistic content, typical approaches are based on the extraction of low-level features. Common features include rhythm, pitch, intensity, etc. We should note that such non-linguistic methods easily provide language-independent models, yet they may be affected by cultural particularities.

Emotion recognition may be used for several reasons. Most popular fields of application include: (a) dynamical marketing, adaptive to the emotional reactions of users (i.e., potential customers) (Ren et al., 2015); (b) smart cars, recognizing the drivers' mood for prevention of accidents due to an unpleasant emotional state (Leng, Lin, & Zanzi, 2007); (c) evaluation of personality of candidates, e.g., during an interview (Lin, Kannappan, & Lau, 2013); (d) evaluation of employees and of user satisfaction in call centers (Petrushin, 1999); (e) enhancing gaming experience by understanding the players' emotional response, etc. (Psaltis et al., 2016). In previous work (Spyrou, Vretos, Pomazanskyi, Asteriadis, & Leligou, 2018) we have applied the assessment of the user affect based on the non-linguistic content of speech for personalization of a non-linear education process. For example, once the affect of a learner was detected to be out of the flow state tending to boredom, the system automatically increased the skill level of learner, while relaxed when she/he was detected in anxiety. In the context of the ENORASI project, the proposed system collects and maps user experiences by geographical region of interest, taking into account the users' emotional state, recognized while interacting with the VUI. This information will be used in analytics aiming to the enhancement of the provided services.

10.4.3 CNN-Based Speech and Emotion Recognition

The advantages of the CNN-based approaches discussed in the context of image analysis in Sect. 10.3 are also valid in the context of speech signal analysis. This motivated us to focus our research toward this direction. CNNs can be exploited visual feature extractors from spectrograms, which are 2D visual representations of the spectral content of the speech signals (Papakostas & Giannakopoulos, 2018). Spectrograms are extracted from fixed-length segments from a given audio sample. The Short-Time Fourier Transform (STFT) is then applied on the original signal. This way, pseudocolored images of spectrograms are generated. For robustness to noise and also for augmenting datasets we add a background sound (e.g., music). Thus, CNNs are trained on the extracted spectrograms. To provide a multilingual approach, the CNN model can be trained using datasets from different languages.

10.4.4 Higher-Level Decision Making

Besides the decision making implemented by ANN approaches, which resembles low-level cognitive processing, the proposed system considers higher-level decision-making approaches to provide feedback and guidance to the user through the VUI. Higher-level cognition can be modeled by artificial cognitive models capable of reasoning within a knowledge space of high-level, semantically relevant concepts. The knowledge about one or more domains can be described by a set of semantic, high-level, interrelated concepts forming a knowledge space. A cognitive model is a computational model capable of simulating human problem-solving and mental task processes within this knowledge space. In that sense, ENORASI investigates high-level cognitive models capable of reasoning based on multiple input concepts related to multiple recognized events (after low-level cognitive processing). The reasoning process results in inference of decisions, e.g., suggesting to the subject which direction to follow, situation assessment, e.g. risk assessment about an alternative, and control, e.g. regulate reaction and generate command for action.

The theory of fuzzy sets provides a sound mathematical framework for uncertainty modeling that has proved its effectiveness in a variety of applications. Fuzzy knowledge-based reasoning methods require that knowledge is represented in the form of rules between higher-level concepts, which can be represented as variables with linguistic values (Zadeh, 1983), e.g., the user interaction about the estimated distance from an obstacle can be based on expressions such as “the distance from the obstacle is small”. The fuzzy cognitive map (FCM) approach can be the basis for enhanced networks for dynamic knowledge representation (Papageorgiou & Salmeron, 2013). An FCM is a fuzzy directed graph with causally interrelated nodes that correspond to the concepts involved in a knowledge domain. It is able to reason through an iterative algorithm updating the values of the graph nodes until a steady state is reached (Papageorgiou & Iakovidis, 2013). This approach is considered for navigation purposes according to the paradigm of Vašcák and Hvizdoš (2016), by also exploiting algorithms coping with the traveling salesman and shortest path problems (Kovács, Iantovics, & Iakovidis, 2018).

10.5 User Requirements

Requirements elicitation is the process of seeking, uncovering, acquiring and elaborating requirements for computer-based systems (Zowghi & Coulin, 2005). As a first step in this process, a literature review was performed for that purpose. A recommended approach to system design is the user-centered design process. This is based on an iterative and continuous update process interacting with the end users, analyzing their feedback and adopting their requirements until the final product is developed (Magnusson, Hedvall, & Caltenco, 2018). The human-centered process is currently well-defined and established as an ISO standard, namely, ISO 9241-210:2010 (Human-centered design for interactive systems) (International Organization for Standardization, 2010). The main axes of the human-centered process are the usability and the user experience. The usability is defined as the ability of the developed system, service or product used by specified users to achieve the defined goals effectively, efficiently and satisfactorily within a certain context of use. On the other hand, user experience is defined as the opinion and perception of the user on the system, service or product after their use (Magnusson et al., 2018).

The user requirements for assistive systems for visually impaired individuals were investigated from several studies in the literature. The most relevant ones, with the assistive system presented in this study, are summarized in Table 10.1. A total of ten studies were considered. Two of the studies investigated (Panchanathan, Black, Rush, & Iyer, 2003; Sosa-Garcia & Odone, 2017) have addressed elicitation of user requirements with respect to CV-based systems for low vision and blind individuals. Individuals with visual impairment have an acute auditory sense; therefore audio-based commands and alerts would make the use of an assistive system easier and more helpful. In the doctoral thesis of Fryer (2013), the effects of the quality of audio descriptions, with respect to the engagement of the visually impaired with the digital media were investigated. The findings of that research support the theory that language can be considered as multimodal, in the sense that it can replace vision within a framework of integrating sensory experience. Another study (Panchanathan et al., 2003) describes the user needs elicited during iCare project. That project was aiming to develop an assistive device that would help visually impaired individuals (mainly students) to ‘pick up a book and read it’, to get information about a person standing in front of them, and to have access to the internet by filtering all the unnecessary information.

Four studies on user requirements and design considerations for cloud/GPS-based information systems for the visually impaired individuals were considered. These include a study addressing an assistive system for urban mobility and transportation by using GPS navigation (Perakovic, Periša, & Prcic, 2015); another study for mobility on urban environments using a robotic guidance system (Hersh & Johnson, 2010); a study investigating various issues regarding outdoor mobility, including outdoor travel frequency, travel independency, different barriers, and shortcomings of GPS (Zeng, 2015); and, a study investigating the user requirements to support tactile mobility (Conradie, de Goedelaan, Mioch, & Saldien, 2014).

Three studies included relevant user requirements for designing a mobile service to enhance learning from cultural heritage. In Alkhafaji et al. (2016), which was focusing on the visually impaired individuals, the results indicated that a multi-service approach at cultural heritage sites should be able to cover services for navigation and directions, to spot nearby cultural heritage places, to receive historical information while touring about the place and the sites and to pre-organize a visit/guided tour. Furthermore, 62% of participants said they would like to customize their mobile application based on their interests. The second of the three studies (Asakawa, Guerreiro, Ahmetovic, Kitani, & Asakawa, 2018), was addressing indoor museum spaces, and it was not focusing to visually impaired individuals; however, some important aspects of museum experience for the visually impaired were mentioned, including purposes (socializing with friends and learning on-site while feeling the atmosphere of the museum), mobility issues, inaccessibility

Table 10.1 Summary of user requirements derived from the literature

#	Requirement	Ref.
1.	Real-time performance for detection/recognition tasks	Sosa-Garcia and Odone (2017)
2.	Ease of use, natural/intuitive user interface, acceptable by a broad user population, including senior citizens	Sosa-Garcia and Odone (2017)
3.	A simple training procedure, potentially scalable to new objects and personalization	Sosa-Garcia and Odone (2017)
4.	Tolerance to viewpoint variations	Sosa-Garcia and Odone (2017)
5.	Tolerance to illumination variations	Sosa-Garcia and Odone (2017)
6.	Tolerance to blur, motion blur, out of focus and occlusions	Sosa-Garcia and Odone (2017)
7.	Accuracy of directions and information	Panchanathan et al. (2003)
8.	Time-efficient access to information	Panchanathan et al. (2003)
9.	Alerts for unexpected events	Panchanathan et al. (2003)
10.	Information to help individuals to tour in an area of interest	Panchanathan et al. (2003)
11.	Audio descriptions of high quality	Fryer (2013)
12.	Automatic creation of return route	Perakovic et al. (2015)
13.	Voice navigation in native languages	Perakovic et al. (2015)
14.	Easy-to-use starting method and configuration	Perakovic et al. (2015)
15.	Use of alternative technologies to GPS for position tracking	Perakovic et al. (2015)
16.	Providing information on location, guidance and navigation	Perakovic et al. (2015)
17.	Providing information on facilities surrounding the user	Perakovic et al. (2015)
18.	Providing information about descending and ascending kerbstone	Perakovic et al. (2015)
19.	Providing information on the system operation	Perakovic et al. (2015)
20.	Providing information of arrival to the destination	Perakovic et al. (2015)
21.	Precise about the movement and location of the user 0.5 (m)	Perakovic et al. (2015)
22.	User-friendliness of the mobile terminal device	Perakovic et al. (2015)
23.	Ability of creating priority information	Perakovic et al. (2015)
24.	Economically affordable solution	Perakovic et al. (2015)
25.	Ability of creating pre-announcement prior to arriving to the destination	Perakovic et al. (2015)
26.	Ability of facility identification	Perakovic et al. (2015)
27.	Selection of the device operation mode offline–online	Perakovic et al. (2015)
28.	Personalization by giving the ability to the users to define their own level of disability	Perakovic et al. (2015)
29.	Minimize the dangers and errors by preventing consequences of incidental or unintentional activity	Perakovic et al. (2015)

(continued)

Table 10.1 (continued)

#	Requirement	Ref.
30.	Sharing information for accompanying contents of surroundings (coffee shops, hotels, hospitals, etc.)	Perakovic et al. (2015)
31.	Compatibility of the device with web applications	Perakovic et al. (2015)
32.	Keyboard as an additional component since visually impaired people are not familiar with touch screens	Perakovic et al. (2015)
33.	Integration with geo-location services with pre-defined SMS messaging	Perakovic et al. (2015)
34.	A multi-function device with GPS for orientation: Location and points of interest, such as 1-m GPS position accuracy, surroundings' description and information and identification of entrances	Hersh and Johnson (2010)
35.	Support and/or emergency: contacting police, ambulance, an emergency center and/or the user's family and giving them the user's location, as well as the provision of help in case the user gets lost	Hersh and Johnson (2010)
36.	A camera for detecting obstacles for also obstacle avoidance (moving and static objects/obstacles' shape, location, moving speed, etc.)	Hersh and Johnson (2010)
37.	Navigation and way-finding, such as finding a street name and safe route	Hersh and Johnson (2010)
38.	Distance and arrival time to the destination as well as the route properties, such as steps up or down, a bridge or a crossing	Hersh and Johnson (2010)
39.	A recording and/or memory function for routes to help the user retrace the route, learn from their mistakes and prepare for future journeys	Hersh and Johnson (2010)
40.	Weather conditions and terrain type notification to support long distance walking and walking in unknown or little known areas	Hersh and Johnson (2010)
41.	Recognizing the color of clothes	Hersh and Johnson (2010)
42.	Discreet and unobtrusive and not attract (undue or unwelcome) attention, including by making unnecessary sounds or noisy operation, or looking exotic, unusual or like medical equipment	Hersh and Johnson (2010)
43.	Attractive and elegant, possibly with a choice of different colors, but in an understated rather than attention grabbing way	Hersh and Johnson (2010)
44.	It should be robust, last a long time and not require maintenance, as well as resistant to damage, pressure, knocks and bumps, water and weather	Hersh and Johnson (2010)
45.	Simple and intuitive to use and look after, including by older people	Hersh and Johnson (2010)
46.	Extending battery life by the device only being powered for steering round obstacles and not for forward motion	Hersh and Johnson (2010)

(continued)

Table 10.1 (continued)

#	Requirement	Ref.
47.	A combination of methods for receiving information from and giving instructions to the robot, though one respondent felt that speech output was the most accessible to all blind and visually impaired people	Hersh and Johnson (2010)
48.	Both a loudspeaker and an earpiece should be available	Hersh and Johnson (2010)
49.	Instructions should be provided by speech and other ways, e.g., a joystick with a scrolling menu and push buttons	Hersh and Johnson (2010)
50.	Speech should be of good quality, sound human not mechanical and be pronounced clearly, with options to change the voice and regulate the volume and rate of delivery	Hersh and Johnson (2010)
51.	A security system to avoid theft, a connection to the user to avoid losing the robot, a manually operated brake, the avoidance of cables and metal parts and that the robot should have knowledge of self-defense	Hersh and Johnson (2010)
52.	An affordable price	Hersh and Johnson (2010)
53.	A USB port and/or wireless connection to update software and/or load data, text and music	Hersh and Johnson (2010)
54.	Software to automatically upgrade on contact with wireless internet but also an internet and/or PC connection to update data and the software	Hersh and Johnson (2010)
55.	The user should be able to move fast with the system, including upstairs and downstairs and in busy situations	Hersh and Johnson (2010)
56.	The design should have few crevices and bends where dirt accumulates and which are difficult to clean. White was considered a color to be avoided for this reason	Hersh and Johnson (2010)
57.	A barcode or RFID reader to read information from barcodes and RFID tags	Hersh and Johnson (2010)
58.	Any speech recognition system used should be of good quality and work well in noisy environments	Hersh and Johnson (2010)
59.	A small display to enable sighted people to read the information, as well as the use of large visual and tactile symbols	Hersh and Johnson (2010)
60.	User choice as to when they received information, particularly spoken information, to avoid irritation to them and other people and attracting attention	Hersh and Johnson (2010)
61.	The system should be re-programmable for the particular user, including accommodating the requirements of users with learning difficulties or other impairments	Hersh and Johnson (2010)
62.	Identification and information of buildings' entrances	Zeng (2015)
63.	Alert for irregular sidewalks	Zeng (2015)
64.	Identification and information of stairs	Zeng (2015)

(continued)

Table 10.1 (continued)

#	Requirement	Ref.
65.	Early alert for obstacles especially in a waist level	Zeng (2015)
66.	Position restore actions when the user gets lost	Zeng (2015)
67.	Roadside holes alert	Zeng (2015)
68.	Information about pedestrian crossings especially in complex forms	Zeng (2015)
69.	Environmental accessibly data	Zeng (2015)
70.	Up-to-date map data	Zeng (2015)
71.	High GPS location accuracy	Zeng (2015)
72.	Strong signal of GPS in urban environment	Zeng (2015)
73.	Notification of uneven floor surfaces such as loose street tiles, puddles or other small holes	Conradie et al. (2014)
74.	The assistive devices should take into account the people with a walking impairment	Conradie et al. (2014)
75.	Systems should reliably provide relevant information when needed, while also considering information accuracy	Conradie et al. (2014)
76.	Designers should also consider providing critical features such as re-location or re-positioning, to allow users to find their way back	Conradie et al. (2014)
77.	Users should be provided with system status information that is critical to use. This may include battery status or current system accuracy	Conradie et al. (2014)
78.	Devices that are used outdoor may need easy ways of recharging batteries, or make use of external batteries	Conradie et al. (2014)
79.	System complexity should also be avoided, to prevent long training times	Conradie et al. (2014)
80.	It should not interfere with other safety relevant interaction mechanisms	Conradie et al. (2014)
81.	Audio should not be the main mode of feedback, especially in situations where users rely heavily on sound to locate and orientate themselves	Conradie et al. (2014)
82.	Alternatives to in-ear earphones may be considered, but critical system information is best communicated via alternative means	Conradie et al. (2014)
83.	The types of obstacles that are communicated to the user should be restricted to those that are unexpected. This is especially important to limit information overload and reduce system complexity	Conradie et al. (2014)
84.	Different contexts may require different types of user interaction. Environments with many obstacles may require different types of notifications (i.e.: more frequent, closer in range)	Conradie et al. (2014)
85.	A balance between the wearing location of both the input sensors and the tactile feedback is needed to ensure the best user experience, while also providing the best results	Conradie et al. (2014)

(continued)

Table 10.1 (continued)

#	Requirement	Ref.
86.	Providing directions and navigation	Alkhafaji et al. (2016)
87.	Identify and give information for nearby cultural heritage places	Alkhafaji et al. (2016)
88.	Find the nearest services	Alkhafaji et al. (2016)
89.	Get historical information while people walk around, and finding out extra information about the sites	Alkhafaji et al. (2016)
90.	Pre-organize visits	Alkhafaji et al. (2016)
91.	Configuration options for personalized customization of the users' mobile app based on their interests	Alkhafaji et al. (2016)
92.	Providing services for cultural heritage information	Alkhafaji et al. (2016)
93.	Operation in parallel with guided tours and in respect to the cultural heritage place so the user will not be fully distracted from tour	Alkhafaji et al. (2016)
94.	Ability to operate offline to avoid poor network quality issues	Alkhafaji et al. (2016)
95.	User-friendly interface and easy operational menu	Alkhafaji et al. (2016)
96.	Weather proof devices in case of open space cultural heritage sites	Alkhafaji et al. (2016)
97.	Detailed audio content is necessary to gain new knowledge	Asakawa et al. (2018)
98.	To listen to the audio contents in front of the artworks, in order to have a similar experience to sighted people	Asakawa et al. (2018)
99.	To listen to human voices as long as they are neutral rather than too emotional	Asakawa et al. (2018)
100.	Concerning the content of the audio descriptions, an introduction/summary, the history, and a detailed visual description of the artwork, followed by detailed descriptions of the technique used should be provided	Asakawa et al. (2018)
101.	To be able to adjust the length of (or skip) the descriptions	Asakawa et al. (2018)
102.	Device functions to be in cooperation with sighted companions	Asakawa et al. (2018)
103.	Identified signs for stairs and toilets	Handa et al. (2010)
104.	Quality of service	Handa et al. (2010)
105.	Cultural information for original artwork	Handa et al. (2010)
106.	Pre-visit information on website and information from museum's website	Handa et al. (2010)
107.	Audio guides with additional information for those who want to get deeper knowledge	Handa et al. (2010)
108.	Assistive system to replace the staff's assistance by providing better quality of interpretation	Handa et al. (2010)

of artworks. The third study (Handa, Dairoku, & Toriyama, 2010) investigated priority needs in terms of museum service accessibility for visually impaired visitors.

10.6 Discussion

The surveys performed in Sects. 10.3–10.5 indicate that current imaging, computer vision, speech, emotion recognition, and decision-making technologies have the potentials to be evolved and integrated into an effective assistive system for the navigation and guidance of the visually impaired individuals. The ENORASI project investigates novel solutions to the challenges involved, aiming to deliver the integrated system described in Sect. 10.2, which should provide enhanced usability and accessibility.

Object detection and recognition are important features in the context of such a system. Object detection approaches, such as Overfeat (Sermanet et al., 2013) and R-CNN (Girshick et al., 2014) are based on two-stage detection approaches, which can be very accurate, yet they are generally computationally demanding. On the other hand, single shot detectors such as YOLO (Redmon et al., 2016) and SSD (Liu, Anguelov, et al., 2016) can provide real-time detection performance, while maintaining reasonable computational requirements; however, they tend not to be very competitive in terms of detection accuracy, as compared with the two-stage detectors. The need for real-time multi-scale obstacle detection for the guidance of the visually impaired increases the demand for the development of robust, lightweight multi-scale object detectors. The state-of-the-art LB-FCN (Diamantis et al., 2019) is considered as a solution toward multi-scale feature extraction and low computational requirements, as compared to conventional networks, such as VGGNet (Simonyan & Zisserman, 2014). Recent deep learning approaches for object recognition, such as GoogLeNet (Szegedy et al., 2015), ResNet (He et al., 2016) and DenseNet (Huang et al., 2017), have provided high object classification accuracies. These architectures require a large number of free parameters which increases their computational complexity, and thus their usage is limited to high-end workstations and servers. Recently, architectures, such as MobileNets (Howard et al., 2017) and ShuffleNets (Zhang, Zhou, et al., 2017), have been proposed aiming to solve this problem by using a fewer free parameters. These networks have been developed specifically to meet the needs of mobile and embedded computing, by making compromises between recognition accuracy and computational complexity. Such compromises may be acceptable in some object recognition applications; however, the guidance of the visually impaired requires real-time performance along with high classification accuracy. The benchmarks performed in Diamantis et al. (2019) suggest that LB-FCN can be considered as a solution to object detection in the proposed system, since it is characterized by smaller computational complexity than relevant architectures, which can be further reduced by utilizing approaches such as the depth-wise separable convolution (Howard et al., 2017). Recognition of multiple objects can be achieved by properly extending LB-FCN for multi-label classification (Vasilakakis et al., 2018).

The estimation of the distance traveled by the user through VO can be considered as a key element of a navigation system for the assistance of the visually impaired (Aladren et al., 2016; Lin, Cheng, et al., 2018; Ramesh et al., 2018; Schwarze et al., 2016; Wang, Katzschmann, et al., 2017; Xiao et al., 2015; Yang, Duarte, & Ganz, 2018). Even though interesting results have been reported in these studies, there are still a lot of challenges toward a reliable travel distance estimation methodology, based exclusively on visual cues. However, recent studies have shown that machine learning models, such as RCNNs (Li et al., 2018; Wang, Clark, et al., 2017), perform well in such tasks. Regarding the above, we opt to incorporate a state-of-the-art stereoscopic RGB-D sensor, namely Intel® RealSense™ D435 with machine learning models, in order to achieve accurate object distance estimation and enhance the performance and robustness of VO irrespectively of the environment, i.e., crowded, outdoor, indoor, etc. Another aspect of the visual measurement domain that could further contextualize the detected obstacles and objects toward enhanced users' experience is that of object/obstacle size measurement, based on visual cues. The research work on this domain is very limited. As reported in the studied literature, the accurate estimation of the size of an object depends on the shape of the object/obstacle, the quality of the image (Hao et al., 2015), the accurate distance estimation from the camera to the object to be measured (Iakovidis, Dimas, et al., 2018) and the accurate detection of the object (Mustafah et al., 2012). Taking into account all of the above, in the proposed system we consider the depth information acquired from the aforementioned RGB-D sensor, and the methodology we proposed in Iakovidis, Dimas, et al. (2018) alongside with the object detection methods discussed in Sect. 10.3.2. Since in the research domain of the visual size estimation there is wide space for improvement, extensions of our visual size estimation method are also investigated, along with the potentials of an ANN-based solution trained to estimate the size of an object/obstacle.

Undoubtedly, emotion recognition is a technology that has attracted the interest of numerous applications in the broader field of human–computer interaction and satellite research areas of computer science. Among these fields, emotion recognition from speech is expected to play an important role, since VUIs are continuously spreading in every part of daily life, i.e., within the home environment, the car, during interaction with computers, mobile phones, phone transactions, etc. It is considered as a slightly invasive approach, compared, e.g., to those using cameras or on-body sensors. The proposed non-linguistic approach uses a CNN, trained using raw speech information encoded as a spectrogram image, having potential to be effective even in cross-language situations. The determination of the users' emotional state during the interaction with the VUI may be used to provide a humanistic modality, able to enhance the process of collecting and mapping user experiences per geographical location visited by the users. In addition, near-natural human–machine interaction is complemented by the use of higher-level inference capabilities, through uncertainty-aware cognitive models, such as FCMs.

The human-centered design process is considered as an important process for setting and addressing the user and design requirements through the various stages of the system's development. For the elicitation of the user requirements for assistive systems focusing on visual impaired people's guidance, ten studies have been studied and in total 108 relevant user requirements have been identified, with respect to relevant CV-based systems, audio descriptions, cloud/GPS-based information systems, and services for learning from cultural heritage and museum accessibility. Most of the user requirements involve audio-based functions, tactile-functions, functions for guidance and description of the surrounding environment, requirements for addressing connectivity issues and design-oriented requirements like battery life and device size.

10.7 Conclusions

This chapter presented the concept of a novel system for computer-assisted navigation and guidance of visually impaired individuals to outdoor environments of cultural interest. The presentation of the system was supported by literature reviews targeted to identify state-of-the-art technological advancements and challenges toward its implementation.

The proposed system is wearable, in the form of smart-glasses, and fully intelligent, in the sense that it integrates components enabling both low and higher-level artificial cognitive processing of multimodal data, including audiovisual data acquired by a stereoscopic depth sensing camera and a microphone, and location data provided by a GPS. Core components that better adapt to the goals of the system include DNN architectures, such as LB-FCN and RNNs, which depending on the task, their design takes into account the tradeoff between the required accuracy and complexity. Features of the system include detection of obstacles, object recognition, emotion-aware voice-based interaction with the user, situation-awareness and decision making upon the users' responses and activities, while providing audio descriptions of the cultural sights of interest.

These capabilities render the system well-beyond the state-of-the-art with respect to: (a) conventional audiovisual aids that have been proposed mainly for navigation of indoor museum navigation, and (b) more general, commercially available solutions for outdoor navigation of the visually impaired. Novel features of the proposed system with respect to the state-of-the-art vision-based assistive technologies include: (a) critical object detection and human–machine interaction based on lightweight CNN architectures running on the MPU, aiming to real-time performance for the users' safety; (b) visual measurement capabilities enhanced with object size estimation; (c) recognition of the users' emotions.

Following the user-requirements analysis performed based on the literature, the most significant challenge is to maximize the system's usability, mainly with respect to:

- Real-time response times;
- Tolerance against various conditions related to the users' environment and activities;
- Natural, user-friendly and efficient human–machine interaction prohibiting cognitive overload;
- Accuracy with respect to navigation even in places where the information provided by the GPS is insufficient;
- Energy autonomy.

The proposed system is being developed in the scope of cultural sights accessibility; however, the vision is to provide an extensible solution that could be ultimately used as an everyday gadget that will improve the quality of life of the visually impaired.

Acknowledgments This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH—CREATE—INNOVATE (project code: T1EDK-02070).

References

Aladren, A., López-Nicolás, G., Puig, L., & Guerrero, J. J. (2016). Navigation assistance for the visually impaired using RGB-D sensor with range expansion. *IEEE Systems Journal, 10*, 922–932.

Alkhafaji, A., Fallahkhair, S., Cocea, M., & Crellin, J. (2016). A survey study to gather requirements for designing a mobile service to enhance learning from cultural heritage. In *European Conference on Technology Enhanced Learning* (pp. 547–550). Cham: Springer.

Anagnostopoulos, C.-N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artificial Intelligence Review, 43*, 155–177.

Asakawa, S., Guerreiro, J., Ahmetovic, D., Kitani, K. M., & Asakawa, C. (2018). The present and future of museum accessibility for people with visual impairments. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 382–384). New York, NY: ACM.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*, 2481–2495.

Baltrusaitis, T., McDuff, D., Banda, N., Mahmoud, M., el Kaliouby, R., Robinson, P., & Picard, R. (2011). Real-time inference of mental states from facial expressions and upper body gestures. In *Proceedings of 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)* (pp. 909–914). Washington, DC: IEEE.

Caraiman, S., Morar, A., Owczarek, M., Burlacu, A., Rzeszotarski, D., Botezatu, N., ... Moldoveanu, A. (2017). Computer vision for the visually impaired: The sound of vision system. In *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)* (pp. 1480–1489). Washington, DC: IEEE.

Conradie, P., Goedelaan, G. K. de, Mioch, T., & Saldien, J. (2014). Blind user requirements to support tactile mobility. In *Tactile Haptic User Interfaces for Tabletops and Tablets (TacTT 2014)* (pp. 48–53).

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine, 18*, 32–80.

Csapó, Á., Wersényi, G., Nagy, H., & Stockman, T. (2015). A survey of assistive technologies and applications for blind users on mobile platforms: A review and foundation for research. *Journal on Multimodal User Interfaces, 9*, 275–286.

Cui, L. (2018). MDSSD: Multi-scale Deconvolutional Single Shot Detector for small objects. arXiv preprint arXiv:1805.07009.

Dai, J., Li, Y., He, K., & Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems* (pp. 379–387).

Diamantis, D., Iakovidis, D. K., & Koulaouzidis, A. (2018). Investigating cross-dataset abnormality detection in endoscopy with a weakly-supervised multiscale convolutional neural network. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 3124–3128). Washington, DC: IEEE.

Diamantis, E. D., Iakovidis, D. K., & Koulaouzidis, A. (2019). Look-behind fully convolutional neural network for computer-aided endoscopy. *Biomedical Signal Processing and Control, 49*, 192–201.

Dimas, G., Spyrou, E., Iakovidis, D. K., & Koulaouzidis, A. (2017). Intelligent visual localization of wireless capsule endoscopes enhanced by color information. *Computers in Biology and Medicine, 89*, 429–440.

Elmannai, W., & Elleithy, K. (2017). Sensor-based assistive devices for visually-impaired people: Current status, challenges, and future directions. *Sensors, 17*, 565.

Erhan, D., Szegedy, C., Toshev, A., & Anguelov, D. (2014). Scalable object detection using deep neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Fang, Z., & Scherer, S. (2015). Real-time onboard 6dof localization of an indoor mav in degraded visual environments using a rgb-d camera. In *2015 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 5253–5259). Washington, DC: IEEE.

Forster, C., Zhang, Z., Gassner, M., Werlberger, M., & Scaramuzza, D. (2017). Svo: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics, 33*, 249–265.

Fryer, L. (2013). Putting it into words: The impact of visual impairment on perception, experience and presence. Doctoral dissertation, Goldsmiths, University of London.

Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., & Berg, A. C. (2017). DSSD: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1440–1448).

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).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (pp. 2672–2680).

Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems, 29*, 1645–1660.

Haag, A., Goronzy, S., Schaich, P., & Williams, J. (2004). Emotion recognition using bio-sensors: First steps towards an automatic system. In *Tutorial and Research Workshop on Affective Dialogue Systems* (pp. 36–48). New York, NY: Springer.

Handa, K., Dairoku, H., & Toriyama, Y. (2010). Investigation of priority needs in terms of museum service accessibility for visually impaired visitors. *British Journal of Visual Impairment, 28*, 221–234.

Hao, M., Yu, H., & Li, D. (2015). The measurement of fish size by machine vision-a review. In *International Conference on Computer and Computing Technologies in Agriculture* (pp. 15–32). Cham: Springer.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 2980–2988). Washington, DC: IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*, 1904–1916.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Held, D., Thrun, S., & Savarese, S. (2016). Learning to track at 100 FPS with deep regression networks. In *European Conference Computer Vision (ECCV).*

Hersh, M. A., & Johnson, M. A. (2010). A robotic guide for blind people. Part 1. A multi-national survey of the attitudes, requirements and preferences of potential end-users. *Applied Bionics and Biomechanics, 7*, 277–288.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*, 359–366.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *CVPR* (p. 3).

Huang, G., Sun, Y., Liu, Z., Sedra, D., & Weinberger, K. Q. (2016). Deep networks with stochastic depth. In *European Conference on Computer Vision* (pp. 646–661). Cham: Springer.

Iakovidis, D. K., Dimas, G., Karargyris, A., Bianchi, F., Ciuti, G., & Koulaouzidis, A. (2018). Deep endoscopic visual measurements. *IEEE Journal of Biomedical and Health Informatics*. https://doi.org/10.1109/JBHI.2018.2853987

Iakovidis, D. K., Georgakopoulos, S. V., Vasilakakis, M., Koulaouzidis, A., & Plagianakos, V. P. (2018). Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification. *IEEE Transactions on Medical Imaging, 37*, 2196–2210.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50× fewer parameters and <0.5 mb model size. arXiv preprint arXiv:1602.07360.

International Organization for Standardization. (2010). ISO 9241-210:2010. https://www.iso.org/standard/52075.html.

Kaur, B., & Bhattacharya, J. (2018). A scene perception system for visually impaired based on object detection and classification using multi-modal DCNN. arXiv preprint arXiv:1805.08798.

Konda, K. R., & Memisevic, R. (2015). Learning visual odometry with a convolutional network. *VISAPP, 1*, 486–490.

Kovács, L., Iantovics, L., & Iakovidis, D. (2018). IntraClusTSP—An incremental intra-cluster refinement heuristic algorithm for symmetric travelling salesman problem. *Symmetry, 10*, 663.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*, 2278–2324.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., … Twitter, W. S. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR* (p. 4).

Leng, H., Lin, Y., & Zanzi, L. (2007). An experimental study on physiological parameters toward driver emotion recognition. In *International Conference on Ergonomics and Health Aspects of Work with Computers* (pp. 237–246). Berlin, Heidelberg: Springer.

Li, R., Wang, S., Long, Z., & Gu, D. (2018). Undeepvo: Monocular visual odometry through unsupervised deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 7286–7291). Washington, DC: IEEE.

Lin, B.-S., Lee, C.-C., & Chiang, P.-Y. (2017). Simple smartphone-based guiding system for visually impaired people. *Sensors, 17*, 1371.

Lin, D. T., Kannappan, A., & Lau, J. N. (2013). The assessment of emotional intelligence among candidates interviewing for general surgery residency. *Journal of Surgical Education, 70*, 514–521.

Lin, S., Cheng, R., Wang, K., & Yang, K. (2018). Visual localizer: Outdoor localization based on convnet descriptor and global optimization for visually impaired pedestrians. *Sensors, 18*, 2476.

Lin, S., Wang, K., Yang, K., & Cheng, R. (2018). KrNet: A kinetic real-time convolutional neural network for navigational assistance. In *International Conference on Computers Helping People with Special Needs* (pp. 55–62). Berlin: Springer.

Lin, T.-Y., Dollár, P., Girshick, R. B., He, K., Hariharan, B., & Belongie, S. J. (2017). Feature pyramid networks for object detection. In *CVPR* (p. 4).

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2018). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2018.2858826

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European Conference on Computer Vision* (pp. 21–37). Cham: Springer.

Liu, Y., Yu, X., Chen, S., & Tang, W. (2016). Object localization and size measurement using networked address event representation imagers. *IEEE Sensors Journal, 16*, 2894–2895.

Luo, W., Li, J., Yang, J., Xu, W., & Zhang, J. (2018). Convolutional sparse autoencoders for image classification. *IEEE Transactions on Neural Networks and Learning Systems, 29*, 3289–3294.

Magnusson, C., Hedvall, P.-O., & Caltenco, H. (2018). Co-designing together with persons with visual impairments. In *Mobility of visually impaired people* (pp. 411–434). Switzerland: Springer.

Maimone, M., Cheng, Y., & Matthies, L. (2007). Two years of visual odometry on the mars exploration rovers. *Journal of Field Robotics, 24*, 169–186.

Mustafah, Y. M., Noor, R., Hasbi, H., & Azma, A. W. (2012). Stereo vision images processing for real-time object distance and size measurements. In *2012 International Conference on Computer and Communication Engineering (ICCCE)* (pp. 659–663). Washington, DC: IEEE.

Nistér, D., Naroditsky, O., & Bergen, J. (2004). Visual odometry. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004 (CVPR 2004)* (pp. I652–I659). Washington, DC: IEEE.

Pan, J., Ferrer, C. C., McGuinness, K., O'Connor, N. E., Torres, J., Sayrol, E., & Giro-i-Nieto, X. (2017). Salgan: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081.

Panchanathan, S., Black, J., Rush, M., & Iyer, V. (2003). iCare-a user centric approach to the development of assistive devices for the blind and visually impaired. In *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence, 2003* (pp. 641–648). Washington, DC: IEEE.

Papageorgiou, E. I., & Iakovidis, D. K. (2013). Intuitionistic fuzzy cognitive maps. *IEEE Transactions on Fuzzy Systems, 21*, 342–354.

Papageorgiou, E. I., & Salmeron, J. L. (2013). A review of fuzzy cognitive maps research during the last decade. *IEEE Transactions on Fuzzy Systems, 21*, 66–79.

Papakostas, M., & Giannakopoulos, T. (2018). Speech-music discrimination using deep visual feature extractors. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2018.05.016

Papakostas, M., Spyrou, E., Giannakopoulos, T., Siantikos, G., Sgouropoulos, D., Mylonas, P., & Makedon, F. (2017). Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. *Computation, 5*, 26.

Perakovic, D., Periša, M., & Prcic, A. B. (2015). Possibilities of applying ICT to improve safe movement of blind and visually impaired persons. In C. Volosencu (Ed.), *Cutting edge research in technologies*. London: IntechOpen.

Petrushin, V. (1999). Emotion in speech: Recognition and application to call centers. In *Proceedings of Artificial Neural Networks in Engineering* (p. 22).

Piana, S., Stagliano, A., Odone, F., Verri, A., & Camurri, A. (2014). Real-time automatic emotion recognition from body gestures. arXiv preprint arXiv:1402.5047.

Poggi, M., & Mattoccia, S. (2016). A wearable mobility aid for the visually impaired based on embedded 3D vision and deep learning. In *2016 IEEE Symposium on Computers and Communication (ISCC)* (pp. 208–213).

Psaltis, A., Kaza, K., Stefanidis, K., Thermos, S., Apostolakis, K. C., Dimitropoulos, K., & Daras, P. (2016). Multimodal affective state recognition in serious games applications. In *2016 IEEE International Conference on Imaging Systems and Techniques (IST)* (pp. 435–439). Washington, DC: IEEE.

Pu, L., Tian, R., Wu, H.-C., & Yan, K. (2016). Novel object-size measurement using the digital camera. In *Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), 2016 IEEE* (pp. 543–548). Washington, DC: IEEE.

Ramesh, K., Nagananda, S., Ramasangu, H., & Deshpande, R. (2018). Real-time localization and navigation in an indoor environment using monocular camera for visually impaired. In *2018 Fifth International Conference on Industrial Engineering and Applications (ICIEA)* (pp. 122–128). Washington, DC: IEEE.

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).

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. arXiv preprint.

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* (pp. 91–99).

Roberts, L. G. (1963). *Machine perception of three-dimensional solids*. Lexington, MA: Massachusetts Institute of Technology (MIT). Lincoln Laboratory.

Schwarze, T., Lauer, M., Schwaab, M., Romanovas, M., Böhm, S., & Jürgensohn, T. (2016). A camera-based mobility aid for visually impaired people. *KI-Künstliche Intelligenz, 30*, 29–36.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.

Shah, N. F. M. N., & Ghazali, M. (2018). A systematic review on digital technology for enhancing user experience in museums. In *International Conference on User Science and Engineering* (pp. 35–46). Singapore: Springer.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sosa-Garcia, J., & Odone, F. (2017). "Hands on" visual recognition for visually impaired users. *ACM Transactions on Accessible Computing (TACCESS), 10*, 8.

Spyrou, E., Vretos, N., Pomazanskyi, A., Asteriadis, S., & Leligou, H. C. (2018). Exploiting IoT technologies for personalized learning. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)* (pp. 1–8). Washington, DC: IEEE.

Suresh, A., Arora, C., Laha, D., Gaba, D., & Bhambri, S. (2017). Intelligent smart glass for visually impaired using deep learning machine vision techniques and robot operating system (ROS). In *International Conference on Robot Intelligence Technology and Applications* (pp. 99–112). Switzerland: Springer.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI* (p. 12).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818–2826).

Tapu, R., Mocanu, B., & Zaharia, T. (2017). DEEP-SEE: Joint object detection, tracking and recognition with application to visually impaired navigational assistance. *Sensors, 17*, 2473.

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern recognition* (4th ed.). Boston: Academic Press.

Tsatsou, D., Pomazanskyi, A., Hortal, E., Spyrou, E., Leligou, H. C., Asteriadis, S., … Daras, P. (2018). Adaptive learning based on affect sensing. In *International Conference on Artificial Intelligence in Education* (pp. 475–479). Switzerland: Springer.

Uijlings, J. R., Van De Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. *International Journal of Computer Vision, 104*, 154–171.

Vašcák, J., & Hvizdoš, J. (2016). Vehicle navigation by fuzzy cognitive maps using sonar and RFID technologies. In *2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI)* (pp. 75–80). Washington, DC: IEEE.

Vasilakakis, M. D., Diamantis, D., Spyrou, E., Koulaouzidis, A., & Iakovidis, D. K. (2018). Weakly supervised multilabel classification for semantic interpretation of endoscopy video frames. *Evolving Systems*, 1–13.

Wang, H., Hu, J., & Deng, W. (2018). Face feature extraction: A complete review. *IEEE Access, 6*, 6001–6039.

Wang, H.-C., Katzschmann, R. K., Teng, S., Araki, B., Giarré, L., & Rus, D. (2017). Enabling independent navigation for visually impaired people through a wearable vision-based feedback system. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 6533–6540). Washington, DC: IEEE.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016). CNN-RNN: A unified framework for multi-label image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, S., Clark, R., Wen, H., & Trigoni, N. (2017). Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 2043–2050). Washington, DC: IEEE.

Wang, X., Gao, L., Song, J., & Shen, H. (2017). Beyond frame-level CNN: Saliency-aware 3-D CNN with LSTM for video action recognition. *IEEE Signal Processing Letters, 24*, 510–514.

WHO: World Health Organization. (2018). Blindness and visual impairement. http://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment.

Xiao, J., Joseph, S. L., Zhang, X., Li, B., Li, X., & Zhang, J. (2015). An assistive navigation framework for the visually impaired. *IEEE Transactions on Human-Machine Systems, 45*, 635–640.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5987–5995). Washington, DC: IEEE.

Yang, K., Wang, K., Bergasa, L. M., Romera, E., Hu, W., Sun, D., … López, E. (2018). Unifying terrain awareness for the visually impaired through real-time semantic segmentation. *Sensors, 18*, 1506.

Yang, K., Wang, K., Zhao, X., Cheng, R., Bai, J., Yang, Y., & Liu, D. (2017). IR stereo realsense: Decreasing minimum range of navigational assistance for visually impaired individuals. *Journal of Ambient Intelligence and Smart Environments, 9*, 743–755.

Yang, Z., Duarte, M. F., & Ganz, A. (2018). A novel crowd-resilient visual localization algorithm via robust PCA background extraction. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1922–1926). Washington, DC: IEEE.

Yu, X., Yang, G., Jones, S., & Saniie, J. (2018). AR marker aided obstacle localization system for assisting visually impaired. In *2018 IEEE International Conference on Electro/Information Technology (EIT)* (pp. 271–276). Washington, DC: IEEE.

Zadeh, L. A. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications, 9*, 149–184.

Zeng, L. (2015). A survey: outdoor mobility experiences by the visually impaired. In *Mensch und Computer 2015–Workshopband*.

Zhang, J., Kaess, M., & Singh, S. (2017). A real-time method for depth enhanced visual odometry. *Autonomous Robots, 41*, 31–43.

Zhang, J., Ong, S., & Nee, A. (2008). Navigation systems for individuals with visual impairment: A survey. In *Proceedings of the Second International Convention on Rehabilitation Engineering & Assistive Technology* (pp. 159–162). Singapore: Singapore Therapeutic, Assistive & Rehabilitative Technologies (START) Centre.

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2017). ShuffleNet: An extremely efficient convolutional neural network for mobile devices. ArXiv e-prints.

Zowghi, D., & Coulin, C. (2005). Requirements elicitation: A survey of techniques, approaches, and tools. In *Engineering and managing software requirements* (pp. 19–46). Berlin, Heidelberg: Springer.