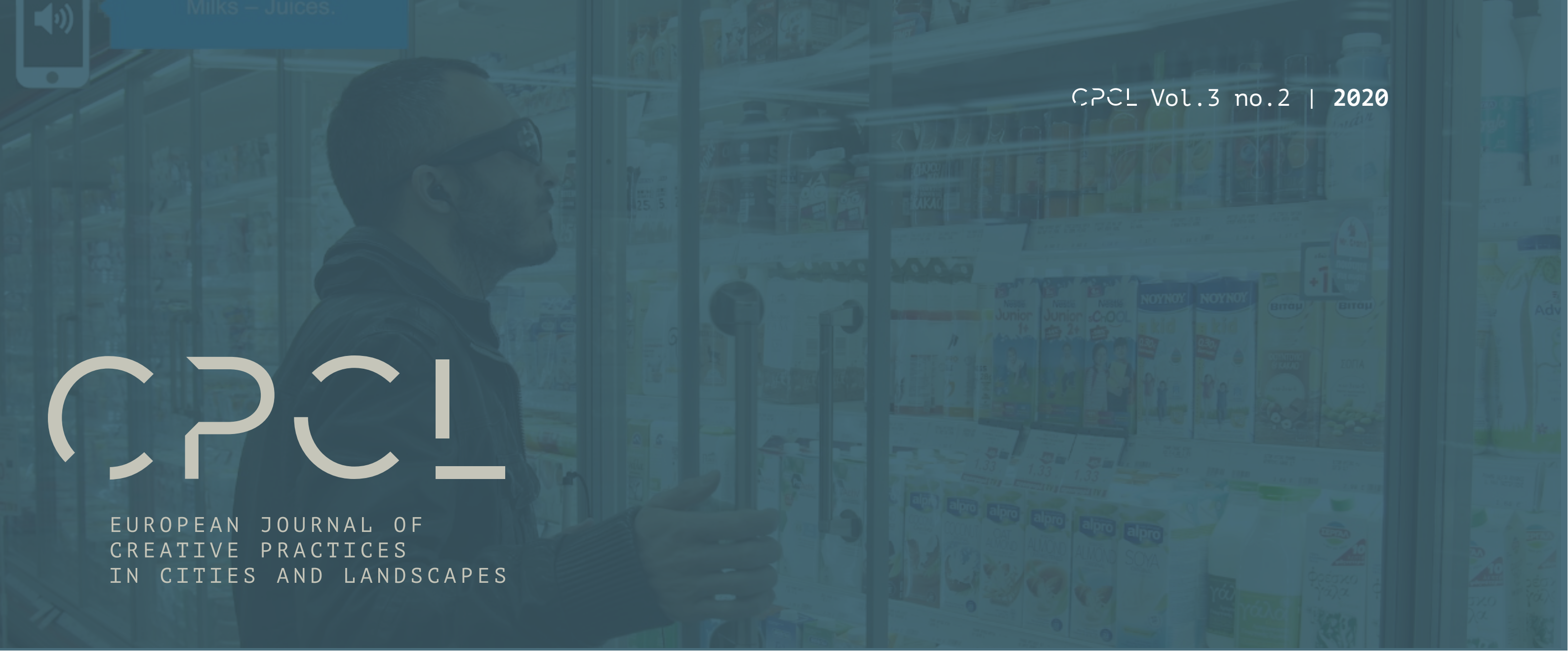


CPCL Vol.3 no.2 | 2020

CPCL

EUROPEAN JOURNAL OF
CREATIVE PRACTICES
IN CITIES AND LANDSCAPES

MAIN SECTION

e-Vision: An AI-powered System for Promoting the Autonomy of Visually Impaired

Panagiotis Migkotzidis — Information Technologies Institute, Greece — Contact: migkotzidis@iti.gr
Fotis Kalaganis — Information Technologies Institute, Greece — Contact: kalaganis@csd.auth.grt
Kostas Georgiadis — Information Technologies Institute, Greece — Contact: kostas.georgiadis@iti.gr
Elisavet Chatzilari — Information Technologies Institute, Greece — Contact: ehatzi@iti.gr
George Pehlivanides — Tetragon S.A., Greece — Contact: g.pehlivanides@gmail.com
Spyros Tsafaras — Tetragon S.A., Greece — Contact: expo@tetragon.gr
Kostas Monastiridis — Tetragon S.A., Greece — Contact: interaction@tetragon.gr
George Martinidis — Aristotle University of Thessaloniki, Greece — Contact: gmart55@yahoo.com
Spiros Nikolopoulos — Information Technologies Institute, Greece — Contact: nikolopo@iti.gr
Ioannis Kompatsiaris — Information Technologies Institute, Greece — Contact: ikom@iti.gr

ACKNOWLEDGEMENTS

European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness

ABSTRACT

Computer vision-based assistive technology for the visually impaired is still a field of ongoing research. Its fundamental scope is to extend the frontiers of visually impaired by means of providing a greater degree of independence and autonomy in their daily living activities. Towards this direction, we present "e-Vision", a hybrid system that couples the convenience and the inherently seamless adoption of an external camera embedded within a pair of eyeglasses with the processing power of modern smartphone devices. The system consists of a pair of eyeglasses integrating a camera and a mobile application that encapsulates computer vision algorithms capable of enhancing several daily living tasks for the visually impaired. The proposed system is a context-aware solution and builds upon three important day-to-day activities: visiting a super-market, going an outdoor walk, and carrying out a work at a public administration building. Going one step further, "e-Vision" also caters for social inclusion by providing social context and enhances overall experience by adopting soundscapes that allow users to perceive selected points of interest in an immersive acoustic way.

KEYWORDS

Computer Vision; Assistive Technologies; Blind

PEER REVIEWED

https://doi.org/10.6092/issn.2612-0496/10707

ISSN 2612-0496

Introduction

Computer vision (CV) is an inextricably connected component, and one of most prominent subfields of Artificial Intelligence (AI), that describes the ability of machines to process and understand visual data. The key concept of CV is to automate the type of tasks the brain's visual processing system, supported by the visual organs (i.e. eyes), typically does. Since its infancy, CV has grown, particularly in the last decade, mainly due to increased data availability and computational power either offered by cloud technologies or by developing dedicated hardware. This has given rise to a number of assistive applications that can replace the human visual system, opting to help the visually impaired in perceiving the world in a similar way to the seeing ones.

The CV-based assistive technology for the blind and visually impaired is still an area under development. It mainly concerns the analysis of images and videos captured by a wearable camera, typically mounted on the chest or head, and provides an egocentric perspective of the world. This point of view is naturally suited to gathering visual information about day-to-day observations and interactions, which in turn can uncover the attention, behavioral structures, and goals of its wielder. The principal objectives of CV-based assistive technology evolve around providing independence and autonomy by enhancing everyday activities of visually impaired. Even though a wide variety of assistive technologies is currently available for the blind, most of them are limited to recognizing obstacles and generic objects without taking into account the context of the activities performed by the user. This context is capable of significantly shaping the functional requirements and consequently enhancing the capabilities of an assistive device.

One of the most notable efforts towards the creation of assistive technologies for the totally blind is the vOICe system[1] which offers the experience of live camera views through image-to-sound renderings and is based on the concept of sensory substitution. With a left to right scanning procedure, images are converted into sound where elevation is associated with pitch while brightness with loudness. From a theoretical neuroscience perspective, this could lead to synthetic vision with actual visual sensations, by taking advantage of the neural plasticity that governs the human brain, through training. Another notable effort concerns the Tyflos system.[2] The Tyflos system consists of camera and Global Positioning System (GPS) sensors, microphones, an audio recording device and

[1] Malika Auvray, Sylvain Hanneton, and J Kevin O'Regan, "Learning to Perceive with a Visuo — Auditory Substitution System: Localisation and Object Recognition with 'The Voice,'" *Perception* 36, no. 3 (March 2007): 416–30, https://doi.org/10.1068/p5631.

[2] Nikolaos Bourbakis et al., "A Multimodal Interaction Scheme between a Blind User and the Tyflos Assistive Prototype," in *2008 20th IEEE International Conference on Tools with Artificial Intelligence* (2008 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Dayton, OH, USA: IEEE, 2008), 487–94, https://doi.org/10.1109/ICTAI.2008.52.

a 2D vibrating vest. A portable computer is used for the purposes of text-to-speech and language processing as well as image analysis. The Tyflos system incorporates a stereoscopic vision module, which is attached to a conventional pair of eyeglasses and is capable of creating a depth map from the surrounding environment. The acquired depth map is converted to a tactile vocabulary that allows the user to perceive his surroundings through a vibratory vest.

More recently, several commercial solutions have been introduced for assisting the visually impaired by exploiting recent advances of computer-vision. These solutions can be categorized into two different types of systems. The first category concerns the smartphone-based systems with the most indicative being seeingAI,[3] Envision[4] and eye-D.[5] These applications take advantage of smartphones' built-in sensors (e.g. camera, accelerometer, etc.) and have recognizing capabilities. More specifically, they allow the user to select from generic categories for recognition, such as reading text, barcodes scanning, detecting people, etc. Then, the recognized instances are narrated to the user though speakers. Besides the smartphone-based category, the second approach concerns systems based on glasses. Prominent examples of this category are OrCam MyEye 2,[6] eSight,[7] NuEyes[8] and Eyesynth.[9] OrCam MyEye 2 is a mobile device with an integrated camera that can be attached to the users' glasses and is capable of recognizing up to 100 custom objects according to user's input (e.g. selected products, people), read text and recognize barcodes. On the other hand, eSight and NuEyes are glass-based devices that work as digital magnifiers, and therefore are only suitable for people with partial visual loss. Finally, Eyesynth is a pair of glasses accompanied by a portable microcomputer that converts the user's 3D surroundings into intuitive sounds that are communicated through cochlear audio and can be used mainly for avoiding obstacles.

In contrast with the aforementioned solutions, the proposed system, namely "e-Vision", is a hybrid approach that couples the natural and seamless adoption provided by an external camera embedded on a pair of glasses with the processing power and the penetration rate of modern smartphone devices. In addition to the system design, e-Vision's main novelty lies in the context-aware design of the application. The structure of the application is built upon specific concepts (i.e. daily-life activities) so as to take advantage of each context accordingly. Consequently, the communication of the system with the user is hassle-free, providing a

[3] https://www.microsoft.com/en-us/ai/seeing-ai, accessed 28 December 2020.

[4] https://www.letsenvision.com/, accessed 28 December 2020.

[5] https://eye-d.in/, accessed 28 December 2020.

[6] https://www.orcam.com/en/myeye2/, accessed 28 December 2020.

[7] https://esighteyewear.com/, accessed 28 December 2020.

[8] https://nueyes.com/, accessed 28 December 2020.

[9] https://eyesynth.com/, accessed 28 December 2020.

pleasant context-aware experience. We should note that the scope of the proposed system is twofold. Apart from increasing the autonomy and independence of the visually impaired, the e-Vision system caters for their social inclusion and aims to promote their overall experience. This is achieved by means of providing social-related information (e.g. emotion analysis of people) and through thoroughly molded soundscapes that allow the user to perceive selected points of interest in a culture-oriented manner.

A survey on the needs of visually impaired people has shown that most of them concern access to information and movement.[10] This is why existing solutions mostly focus on the identification of obstacles. Navigation is indeed a vital need, but only a rudimentary one. This is supported by the fact that the most common questions that they pose to their sighted peers concern the identification and description of objects around them.[11]

The need to identify objects serves practical purposes, since, in the aforementioned survey, the majority of the visually impaired (60%) mentioned shopping as the everyday activity in which they require the most help. This is one of the key activities that e-Vision is supporting. A great part of e-Vision's -and the present study's- originality is that the system is designed to cover such needs that are not addressed by existing systems.

1. e-Vision overview

The principal objective of e-Vision is to promote the autonomy and independence of people with visual impairment. A crucial component to achieve this objective concerns the creation of a mobile application capable of enhancing their daily activities to an optimal level. In order to achieve this ultimate goal, the aforementioned mobile application passes inevitably through the employment of recent technological advancements in the fields of CV and computational intelligence. Considering not only the capabilities, but also the limitations, of the existing technological reality it becomes evident that the development of a universal application suitable to cover all the needs of visually impaired is utopic. Therefore, the development of e-Vision is based on the following assumption; the application should take into account the contextual information of an activity. Therefore, the development of e-Vision is tailored to support three major daily life activities: a) shopping in super-markets, b) going an outdoor walk and c) visiting a public administration building to carry out some bureaucratic task.

[10] D. Gold and H. Simson, "Identifying the Needs of People in Canada Who Are Blind or Visually Impaired: Preliminary Results of a Nation-Wide Study," *International Congress Series* 1282 (September 2005): 139–42, https://doi.org/10.1016/j.ics.2005.05.055.

[11] Erin Brady et al., "Visual Challenges in the Everyday Lives of Blind People," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13: CHI Conference on Human Factors in Computing Systems, Paris France: ACM, 2013), 2117–26, https://doi.org/10.1145/2470654.2481291.

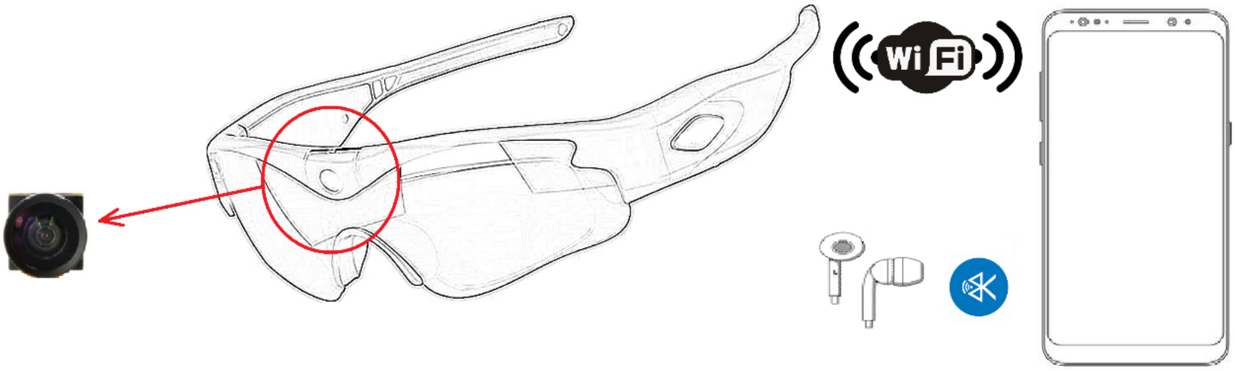


FIG. 1 The main hardware components of the e-Vision system. A pair of glasses with an integrated camera, a smartphone and a pair of earphones. The bidirectional communication between the camera and the smartphone is performed through Wi-Fi.

To achieve efficiency and effectively support the visually impaired, the e-Vision system consists of two hardware components. The first component is a pair of Wi-Fi enabled camera-glasses. More specifically, a wireless camera is integrated in a typical pair of glasses that can wirelessly transmit captured images or video in real-time through Wi-Fi protocol. The second component is a smartphone device, running either iOS or Android operating system, which hosts the e-Vision application and consequently realizes the CV components. The system's feedback is provided to the user by means of narration though the auditory pathway using conventional earphones, a crucial component capable of ensuring the essential seclusion of the feedback. The combination of the aforementioned hardware components ensures both freedom in movement and sufficient computational power in order to realize the essential computer-vision functionalities [Fig. 1].

The e-Vision mobile application[12] has been developed with respect to the SOLID object-oriented programming principles.[13] In this context, we have developed an application that is more understandable, flexible and maintainable. The modular nature of e-Vision allows third-party programmers to replace existing software modules and hardware components, with more suitable options, according to their individual needs. In this way, the e-Vision system will be able to adapt, as the computer-vision technology evolves, in a seamless manner.

1.1 Blind-friendly user interface

The design of an application for the fully and partially blind is a demanding process that remains challenging for both the users and the designers since most of the conventional mobile applications are not addressed to the visually impaired. The literature is characterized by a variety of efforts

[12] The code is available at https://github.com/e-vision-project/vision-API, accessed 28 December 2020.

[13] Robert C. Martin, "Design Principles and Design Patterns," *Object Mentor* 1, no. 34 (2000).

that aim to establish a common framework for designing applications that target the visually impaired. Alonso et al., introduced a set of rules that should overrule applications targeting the visually impaired and suggested a variety of prototypical user models.[14] In 2012, Sierra & De Togores revised structural tools and design elements specifically tailored to satisfy the needs of the visually impaired.[15] More recently, Olofsson studied the design process of applications for the visually impaired by conducting interviews with field experts and visually impaired people. The outcome of her research[16] indicated that the visually impaired prefer multimodal designs that combine graphical, auditory and tactile elements.

By taking into account the existing literature and in close collaboration with the "Center for Education and Rehabilitation for the Blind – CERB"[17] a preliminary research was conducted that uncovered the structural design elements and the essential principles that should be followed in order to develop a blind-friendly user interface (UI) for the e-Vision system. This research included not only questionnaires but also actual interaction with design mockups. With respect to the outcome of the aforementioned research, the developed interface abides to the following rules: a) only essential feedback should be provided to user, b) simplified design elements and patterns should be followed, c) exploitation of special gestures for providing feedback (such as the shake gesture), d) employment of light and high contrast colors, e) avoidance of auditory information overload and f) parameterization of the system's narrative feedback (e.g. text-to-speech narration speed and verbosity).

To this end, the e-Vision system employs a blind-friendly gesture control that emphasizes on simplicity and efficiency. Since e-Vision is addressed to the visually impaired, who typically use a white cane for navigation, the UI is oriented towards one-hand usage. Moreover, the UI has been tailored accordingly in an effort to enable onscreen location invariant interaction. Therefore, the employed gestures could be used without taking into account the location of the screen where gestures take place, without any loss of control precision. Finally, in order to avoid unintended interactions, the smartphone's screen is isolated during the e-Vision usage allowing control only over the introduced system's elements. We note that the assistive technologies of modern smartphone-based operating systems played a crucial role for the design process of e-Vision's UI. By employing

[14] Fernando Alonso et al., "User-Interface Modelling for Blind Users," in *Computers Helping People with Special Needs*, ed. Klaus Miesenberger et al., vol. 5105, Lecture Notes in Computer Science (Berlin, Heidelberg: Springer Berlin Heidelberg, 2008), 789–96, https://doi.org/10.1007/978-3-540-70540-6_117.

[15] Javier Sánchez Sierra and Joaquín Selva Roca de Togores, "Designing Mobile Apps for Visually Impaired and Blind Users: Using Touch Screen Based Mobile Devices: IPhone/IPad," *ACHI 2012: The Fifth International Conference on Advances in Computer-Human Interactions*, 2012, 47–52.

[16] Stina Olofsson, "Designing Interfaces for the Visually Impaired : Contextual Information and Analysis of User Needs" (Umeå University, 2018).

[17] http://www.keat.gr/index.php/en/, accessed 28 December 2020.

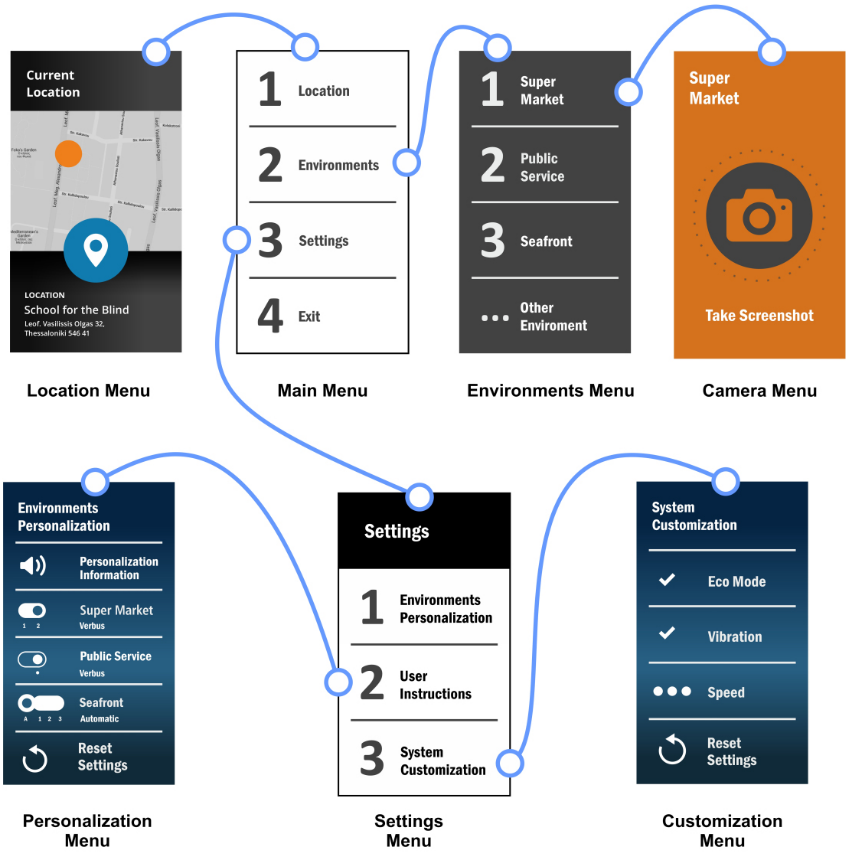


FIG. 2 Illustration of the user interface of the e-Vision application. Blue lines indicate the interface interconnections and transitions

common gestures of widespread assistive technologies, e-Vision ensures familiarity and a steep learning curve. More specifically, is it the following gestures that realize e-Vision's gesture control:

- *Double tap/ Swipe right (one finger)*: Option selection.
- *Swipe down (one finger)*: Move to the next option of the current menu screen.
- *Swipe up (one finger)*: Move to the previous option of the current menu screen.
- *Swipe left (one finger)*: Move to the previous menu screen.
- *Swipe left (two fingers)*: Move to the initial menu screen.
- *Swipe up (two fingers)*: Narrating all the options of the current menu screen.
- Touch and hold: Narrating the current option.

Having in mind the target group of the e-Vision system, the UI design process is inspired by the Voice User Interface paradigm. Therefore, the application narrates each design element, one at a time, in a sequential manner. The followed design approach differs from typical Graphical User Interfaces where users are able to scan the whole interface design and instantly develop a universal understanding, even on the invisible aspects, of the UI's structure. The visually impaired users have the possibility to navigate in a top-to-bottom approach and get auditory feedback both on the targeted option and the corresponding type of element (e.g. button, slider, etc.). Each auditory feedback consists of a) narration that describes the targeted option, b) a distinctive alert sound that provides information regarding the success or failure of an action and c) a distinctive vibration pattern that provides tactile feedback for specific actions. Figure 2, demonstrates a detailed diagram of e-vision's UI and the interconnections between its various parts **[Fig. 2]**.

2. Computer-vision modules

The scientific area of Deep Learning (DL) has recently entered the picture of CV. Although DL's fundamentals have been formulated several decades ago, its rapid growth has been observed only recently mainly due to the abundance of available data nowadays, a prerequisite for any DL model in order to achieve sufficiently high performance. The increased availability of training data can be attributed to the continuous digitization of our society (e.g. mobile phones, social media, etc). The CV modules of e-Vision are built on top of modern DL technologies specifically selected to operate on mobile devices.

The mobile application of e-Vision is developed in a modular manner that allows its seamless modification and extension. Although each module operates independently from the others it is only their combination that can efficiently support the targeted daily activities. The employed modules (image classification, object recognition, facial landmarks and emotion recognition, and optical character recognition) were selected so that each one can complement another towards a common goal, an assistive system for the visually impaired.

2.1 Image classification

Image classification is the process of taking an image as input and classifying it according to its visual content. As an example, an image classification algorithm may be designed to distinguish if an image contains a human figure, a car or something else. The output of such an algorithm is either a class (e.g. "car") or a probability that the input is of a particular class ("there's a 90% chance that this input is a car"). While such a task is trivial for human beings, robust image classification is still an ongoing research topic in CV applications. However, the DL field has significantly shaped the field of CV since it is able to uncover complicated structures in high-dimensional data.

In the context of e-Vision, we employ DL architectures that are specifically tailored to operate in mobile devices. More specifically, we take advantage of MobileNetV2,[18] an architecture which has limited requirements in terms of computational power without significantly sacrificing classification performance. The employed architecture was trained on Imagenet,[19] an image database with a total of one million images and one thousand visual categories. In order to transfer to new concepts, the MobileNetV2 pre-trained model was used in order to extract high-level features and structures from images, while the final classification was performed by linear Support Vector Machines (SVMs), which are sophisticated machine learning algorithms capable of achieving high performance without requiring an excessive amount of data.

2.2 Object recognition

Object detection is a computer technology related to CV and image processing that deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos. In contrast to image classification, object detection outputs multiple classes for an input image as well as a set of boxes which bound the detected objects and provide information about their position within the image.

With the advent of DL and its takeover of the CV field, the object detection algorithms have managed to achieve, in specific tasks, superhuman performance. Current object detectors can be divided into two categories: a) Networks that separately perform the tasks of determining the location of objects and their classification and b) networks which predict bounding boxes and class scores jointly in a single step. The second category is characterized by simplicity, making it an appropriate choice for deployment in mobile devices. Tiny YOLO[20] is among the most notable architectures of this category and was the one that was employed in the e-Vision case. More specifically, it is a single-stage architecture that goes straight from image pixels to bounding box coordinates and class probabilities and runs in real-time as it can achieve more than 10 fps on modern smartphone devices. The model that was used for the need of e-Vision was trained on the Open Images Dataset[21] which contains a total of 16M bounding boxes for 600 object classes on 1.9M images, making it the

[18] Zheng Qin et al., “FD-MobileNet: Improved MobileNet with a Fast Downsampling Strategy,” in *2018 25th IEEE International Conference on Image Processing (ICIP)* (IEEE, 2018), 1363–67.

[19] Jia Deng et al., “ImageNet: A Large-Scale Hierarchical Image Database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), Miami, FL: IEEE, 2009), 248–55, https://doi.org/10.1109/CVPR.2009.5206848.

[20] Joseph Redmon et al., “You Only Look Once: Unified, Real-Time Object Detection,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) (2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, 2016), 779–88, https://doi.org/10.1109/CVPR.2016.91.

[21] Alina Kuznetsova et al., “The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale,” *International Journal of Computer Vision* 128, no. 7 (July 2020): 1956–81, https://doi.org/10.1007/s11263-020-01316-z.

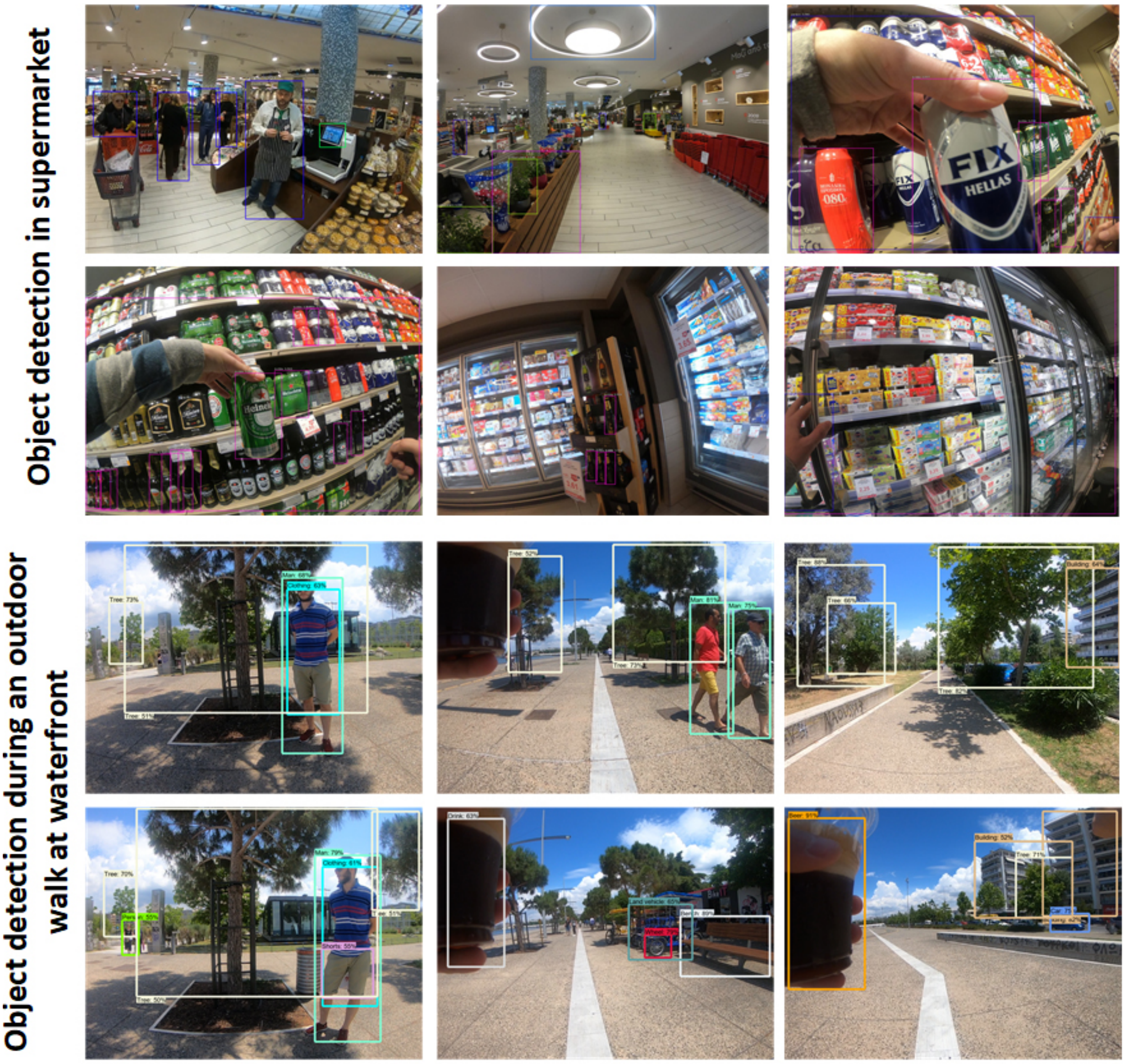

Object detection in supermarket
Object detection during an outdoor walk at waterfront

FIG. 3 Object recognition in exemplar images (egocentric viewpoint) during a visit in supermarket and an outdoor walk.

largest existing dataset with object location annotations. Figure 3 demonstrates the employed object detection model in photos from a visit to the supermarket and an actual outdoor walk [Fig. 3].

2.3 Facial landmarks and emotion recognition

One of the most integral parts of the face recognition pipeline is the verification of the existence of a face in a provided image. This can be achieved through a series of steps, referred to as face detection process, that map facial landmarks (i.e. parts of faces, like eyes or mouth), with their combination providing the prediction. e-Vision exploits the recent advances in one shot learning models, where the training process requires a limited number of faces (a few or even one –in contrast to the typical DNN models–) [Fig. 4].

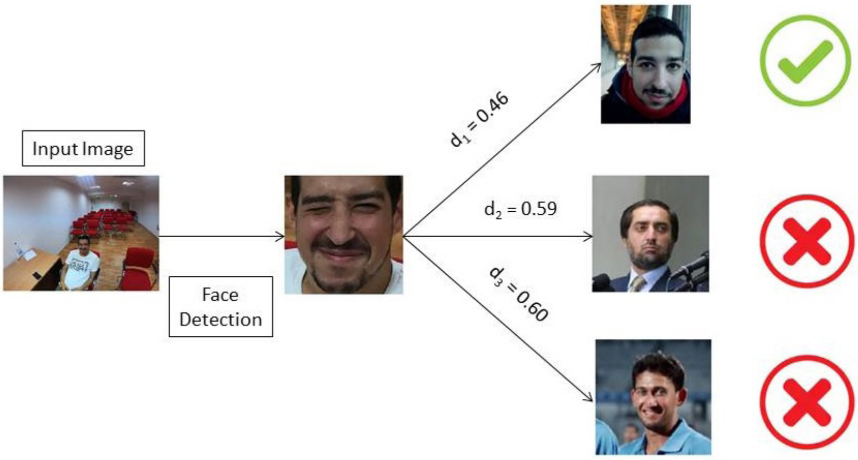
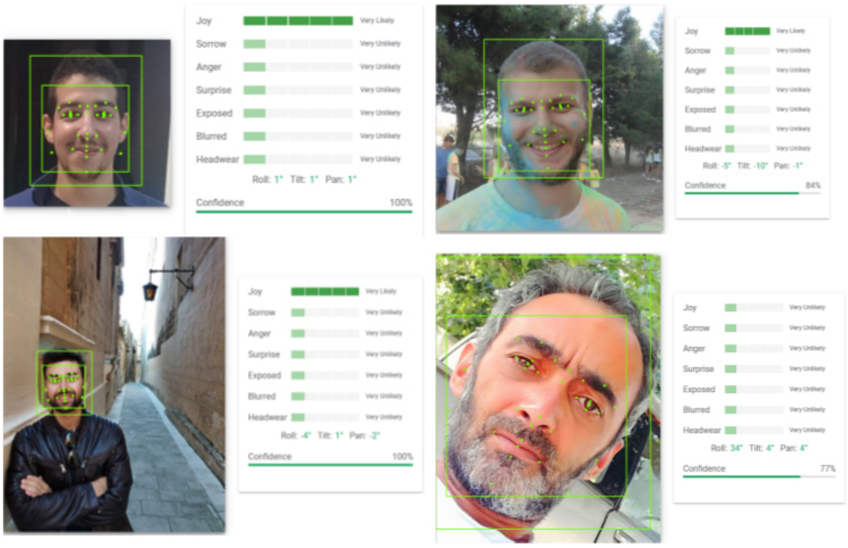


FIG. 4 Example usage of the employed face-recognition algorithm.



Joy		Very Likely
Sorrow		Very Unlikely
Anger		Very Unlikely
Surprise		Very Unlikely
Exposed		Very Unlikely
Blurred		Very Unlikely
Headwear		Very Unlikely

Roll: 1° Tilt: 1° Pan: 1°

Confidence 100%

Joy		Very Likely
Sorrow		Very Unlikely
Anger		Very Unlikely
Surprise		Very Unlikely
Exposed		Very Unlikely
Blurred		Very Unlikely
Headwear		Very Unlikely

Roll: -5° Tilt: -10° Pan: -1°

Confidence 84%

Joy		Very Likely
Sorrow		Very Unlikely
Anger		Very Unlikely
Surprise		Very Unlikely
Exposed		Very Unlikely
Blurred		Very Unlikely
Headwear		Very Unlikely

Roll: -4° Tilt: 1° Pan: -2°

Confidence 100%

Joy		Very Unlikely
Sorrow		Very Unlikely
Anger		Very Unlikely
Surprise		Very Unlikely
Exposed		Very Unlikely
Blurred		Very Unlikely
Headwear		Very Unlikely

Roll: 34° Tilt: 4° Pan: 4°

Confidence 77%

FIG. 5 Illustrating facial landmark detection and emotion recognition by the employed CV model

In this scenario, the model requires three face images that are provided to three identical CNNs that formulate the corresponding embeddings. Then the corresponding pairwise Euclidean distances calculated using a triplet loss function. Figure 4 illustrates an example where the employed tool detects the face in the image and then returns three matching images, the correct photo (the individual's original photo) and two false positive images from the dataset. It is evident that the minimum distance is reached when the correct image is identified, with a significant difference compared to the two false positive cases [Fig. 5].

More recently, DL models and more specifically CNNs have been employed for the detection of facial landmarks towards emotion recognition with remarkable success. The majority of the models can be categorized as pure-learning[22] or hybrid-learning methods,[23] with the former providing a prompt identification of the facial landmarks while the latter combine DL techniques with projection models to cast a prediction that have proven to perform better to a larger variety of expressions. Therefore, for the purposes of this work where higher degrees of freedom are crucial, we employ hybrid-learning methods[24.] An example use of the selected approach is depicted in Figure 5, where at first the facial landmarks of four individuals are identified and then a prediction regarding their emotional state is provided.

2.4 Optical Character Recognition

Optical Character Recognition (OCR) is the mechanism that converts images that contain text (handwritten or printed) to text that can be interpreted by a computer. Recent approaches in the domain of OCR employ DL models and more specifically Recurrent Neural Networks (RNNs) or CNNs with the most prominent paradigms being the ones that also encapsulate feedback connections like the long short-term memory (LSTM) architectures.[25]

The employment of DL architectures can be readily identified in the field's pioneers like Tessaract,[26] where the system's software is built upon LSTM architectures. OCR systems are classified into the ones that perform OCR when a document is scanned and the ones applied on images. In the first case OCR is applied upon scanning a document in order to convert it to a digital file[27] (e.g. PDF) while in the latter the OCR mechanism is enabled when an image is captured. The needs of the developed tool, where users will provide photos to the system, impose the use of software of the second category **[Fig. 6]**.

Considering the functional requirements of e-Vision, the selected software is Google's Vision API, a CNN based OCR software that outperforms other competitors, providing the lowest false detection rate per character.[28] Besides its superiority against competitors, API Vision provides online features of extreme importance for the developed system. Example uses

[22] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep Convolutional Network Cascade for Facial Point Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, 3476–83.

[23] Amin Jourabloo and Xiaoming Liu, "Pose-Invariant Face Alignment via CNN-Based Dense 3D Model Fitting," *International Journal of Computer Vision* 124, no. 2 (September 2017): 187–203, https://doi.org/10.1007/s11263-017-1012-z.

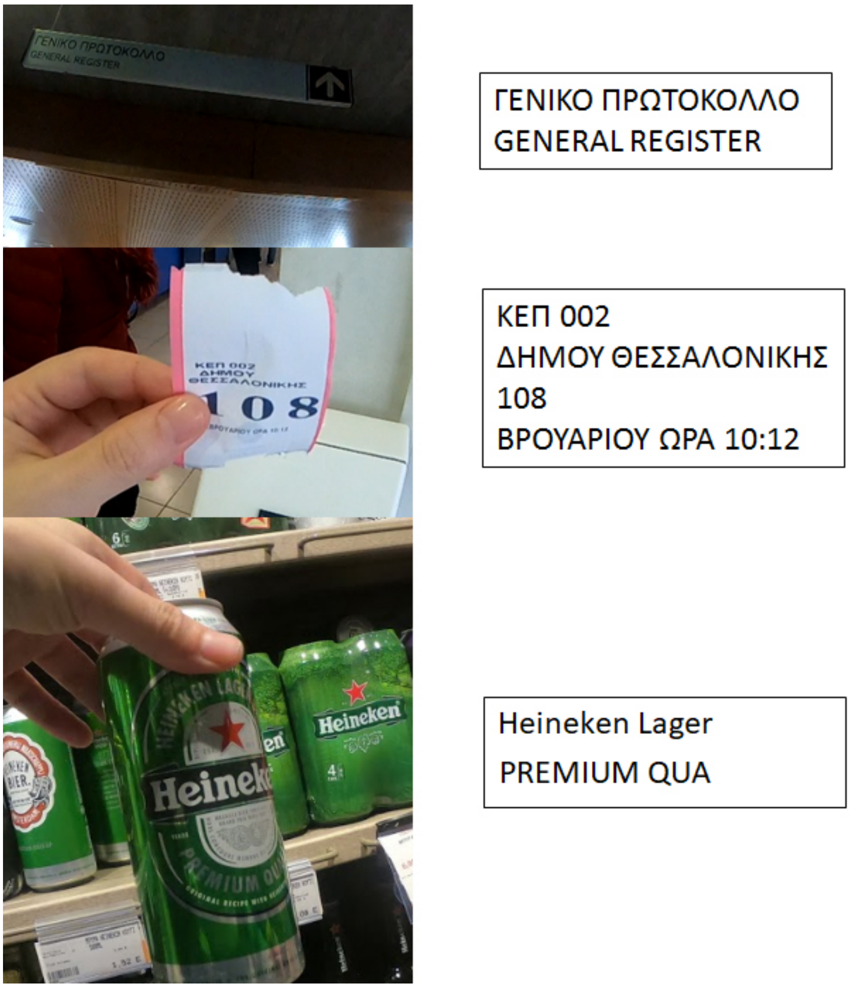
[24] https://cloud.google.com/vision, accessed 28 December 2020.

[25] Christian Bartz, Haojin Yang, and Christoph Meinel, "STN-OCR: A Single Neural Network for Text Detection and Text Recognition," *ArXiv*, 2017.

[26] https://github.com/tesseract-ocr, accessed 28 December 2020.

[27] https://www.abbyy.com/en-ee/finereader/, accessed 28 December 2020.

[28] Jake Walker, Yasuhisa Fujii, and Ashok C. Popat, "A Web-Based OCR Service for Documents," in *Proceedings of the 13th IAPR International Workshop on Document Analysis Systems* (DAS), vol. 1 (Vienna, Austria, 2018).



ΓΕΝΙΚΟ ΠΡΩΤΟΚΟΛΛΟ
GENERAL REGISTER

ΚΕΠ 002
ΔΗΜΟΥ ΘΕΣΣΑΛΟΝΙΚΗΣ
108
ΒΡΟΥΑΡΙΟΥ ΩΡΑ 10:12

Heineken Lager
PREMIUM QUA

FIG. 6 Example usage of the OCR module on actual images (egocentric viewpoint) from a supermarket (bottom image) and a public administration building. Input images on the left and the identified text on the right (mostly Greek characters).

of the selected OCR model applied to various images containing text can be seen in Figure 6. It is evident that in most cases the system recognizes the entirety of the provided text.

3. Supported activities

e-Vision is structured towards supporting three distinct daily activities for the visually impaired. Each one of the supported activities has its own functional requirements and therefore, the e-Vision system is designed to take into account the particular context that accompanies each activity. For each supported activity (i.e. supermarket visit, public administration visit and outdoor walk), several CV modules are employed in a complementary manner. We note that each of these modules is used for different reasons in each supported case. As an indicative example, we present the case of optical character recognition which is either used for document reading in the public administration case or for product identification in the case of the supermarket. In the following sections, we present the operation of the e-Vision system for each of the supported daily activities.

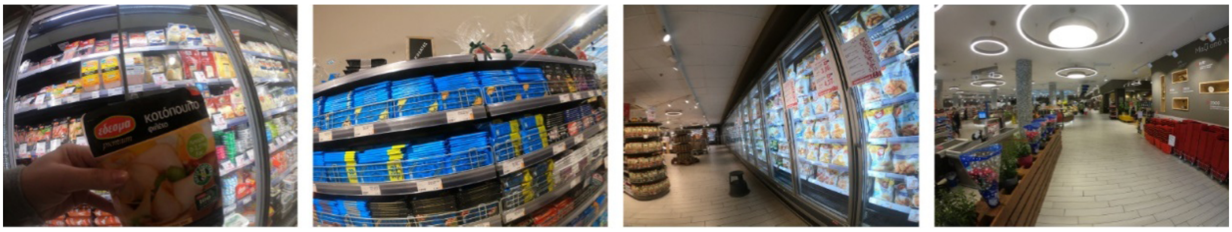


FIG. 7 Egocentric images depicting the classes in the supermarket activity namely (left to right): a) product, b) shelf, c) trail and d) other, from an actual visit to a supermarket by a visually impaired person.

3.1 Supermarket

The first use case scenario provides the visually impaired with the ability to visit a supermarket and complete a session of grocery shopping. Users are able to operate the system simply by (double) tapping the screen of their cell phone whenever they require information. Tapping the screen triggers the Wi-Fi camera to capture a first-person shot and enables a series of actions that will be completed with the system providing relevant information to the users in two levels of abstraction. The first level consists of a general description of what is in front of the user while the second overlays a more elaborate description.

The first level will facilitate navigation in the supermarket, an issue of paramount importance for the visually impaired, providing only essential information to the user while ensuring low levels of frustration. In this direction, the proposed system will classify any given image in one of the four basic concepts/categories: a) product, b) shelf, c) trail and d) other/unrecognized. In order for the system to provide accurate predictions, a combination of DL architectures and SVMs were opted, with the first being opted as feature extractors, while the latter for casting a prediction as described in section 3.1. Example uses for this case scenario, where a visually impaired person visited a supermarket, can be seen in the following figure **[Fig. 7]**.

The second level will exploit the information provided by the first and will in turn identify the specific product, shelf or trail. As a result, users will have access to detailed descriptions about the product they are holding or the trail/shelf they are looking at. This will help the users to semantically navigate in a supermarket and allow them to easily and accurately do their grocery shopping in an autonomous way. The steps required to reach to the aforementioned descriptions, include OCR mechanisms followed by a matching mechanism on a product database provided by Masoutis,[29] one of the largest Greek supermarket corporations. More specifically, OCR extracts text, arising from the product packages (e.g. brand, product description) from the provided image. A search for the identified word(s) is then performed in Masoutis database in order to determine the exact

[29] www.masoutis.gr, accessed 28 December 2020.

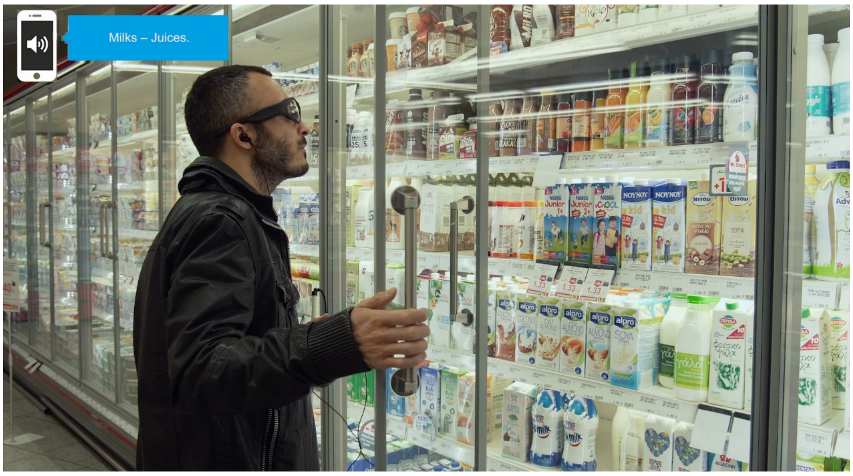


FIG. 8 Usage of the e-Vision application during an actual supermarket visit by a visually impaired person. The blue textbox contains the message being communicated to the user.

product, self or trail category. A more detailed technical description of the supermarket case is available by Kostas Georgiadis et al.[30] [Fig. 8]

3.2 Public administration

Carrying out a task in a public administration building (e.g. getting a birth certificate) is extremely demanding for visually impaired people. Although people with all kinds of impairments are given priority and help by employees in such places, this causes them unease. Therefore, the increase of their autonomy and independence has a positive effect on their mental health.[31] Towards this direction, e-Vision supports several features specially tailored for the public administration case. By taking advantage of the object detection module, the e-Vision system is capable of notifying the user about the existence of a ticket dispenser (e.g. take-a-number system) as well as other useful objects including chairs, desks and people. To increase the social aspect of this activity, by exploiting the face and emotion recognition module, the e-Vision system provides social context during a conversation or a transaction (e.g. "You are facing a happy man"). Finally, the most important feature of this case is realized by the optical character recognition module where the system can read documents and signs to users. All the feedback is provided to the user by means of narration through the auditory pathway **[Fig. 9]**.

[30] Kostas Georgiadis et al., "A Computer Vision System Supporting Blind People - The Supermarket Case," in *Computer Vision Systems*, ed. Dimitrios Tzovaras et al., vol. 11754, Lecture Notes in Computer Science (Cham: Springer International Publishing, 2019), 305–15, https://doi.org/10.1007/978-3-030-34995-0_28.

[31] Daniela Mirandola et al., "Psychological Well-Being and Quality of Life in Visually Impaired Baseball Players: An Italian National Survey," ed. Stefano Federici, *PLoS One* 14, no. 6 (2019): e0218124, https://doi.org/10.1371/journal.pone.0218124.

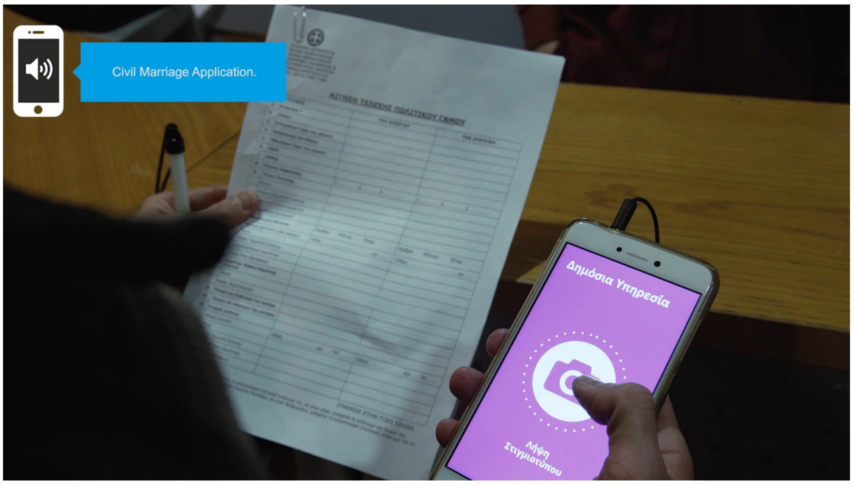


FIG. 9 Usage of the e-Vision application during an actual visit at a public administration by a visually impaired person. The blue textbox contains the message being communicated to the user.

3.3 Outdoor walk

The third of the targeted activities of e-Vision is an outdoor walk. Here the system serves a variety of purposes ranging from an immersive cultural experience through specifically designed soundscapes to user's friends identification and auditory portraiture of the user's surrounding. By exploiting the face recognition advances in CV, the system will be able to identify detected faces and whether they belong to the user's social environment. To enable this feature, each user must provide portraits of his friends, though a specifically designed interface, that will serve as the baseline for the face recognition module. The identified faces are communicated to the user through the auditory pathway (i.e. "Your friend John is approaching."). Moreover, in an effort to provide an auditory depiction of the surroundings, we take advantage of the object detection module. The user's surroundings are processed in order to extract the depicted set of identified objects. Then the identified objects are converted into an elegant narration (e.g. "A bench on your right.") that is provided to the user using text-to-speech technologies. In an effort to avoid auditory -and consequently cognitive- overload, only the essential feedback is provided according to the user's individual preferences, by means of narration verbosity and feedback frequency, through the settings menu of the e-Vision application. To further enhance the system's parameterization and personalization, an on-demand feedback option is also provided where the user requests feedback if desired.

3.3.1 Soundscapes

During an outdoor walk, visually impaired people use their hearing to perceive the surroundings and get information concerning a wide variety of events (e.g. task-related sounds such as running or bike riding and

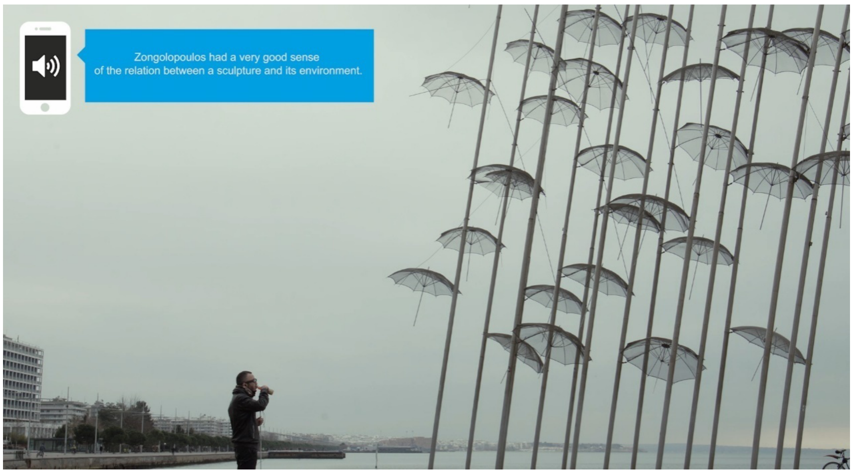

Zongolopoulos had a very good sense
of the relation between a sculpture and its environment.

FIG. 10 Usage of the e-Vision application (soundscape case), during an outdoor walk, near the "Zongolopoulos' umbrellas" monument, by a visually impaired person. The blue textbox contains a sample of the message that is communicated to the user.

natural sounds such as a dog barking, birds tweeting wind blowing etc). Using the e-Vision system, the visually impaired should be able to receive -again through the auditory pathway and in the form of sounds- information about their surroundings, but this should be aligned with the aforementioned natural sounds and should not interfere with their perceptual processes.

In the case of e-Vision, the concept of soundscape is not confined to recording and describing the sounds that shape users' environment since these sounds will, naturally and in real-time, end up in the users' ears as they walk. In the case of e-Vision, the concept of soundscape refers to a set of high-quality and edited audio clips that correspond to selected points of cultural interest such as monuments, museums etc. When users approach a selected point of interest (identified by the smartphone's GPS sensor), they will be alerted with a brief and distinctive tone. If they like, they will be able to hear one or more sound clips that will constitute the corresponding soundscape. Moreover, the users will be able to switch to another audio clip (if one exists) as well as stop the process at any time. By employing such practices, the users will not be isolated from their surroundings, but will be able to also monitor the natural sounds of their environment [Fig. 10].

Conclusion

Despite the recent advances in the field of assistive technologies for visually impaired people, existing practices are limited to providing generic information regarding objects and identifying obstacles. In this study, we introduced the e-Vision system which aims to provide information regarding objects, faces and obstacles in contextualized manner amenable to semantic interpretation. The scope of e-Vision is to take advantage of recent advances in the field of CV in order to assist people with visual impairment and provide them the ability to identify objects and persons at a semantic level. Indeed, developments in CV have brought them closer to the efficiency levels of human vision even without the need for high computational power, enabling their operation on even less sophisticated devices, such as smartphones. The development of CV algorithms complements the wide availability of visual content that is essential for their training, as the ever-expanding streamlining of society is accompanied by large-scale digital content production and sharing. The described functionalities are framed within an interface tailored to visually impaired people that allows the control of parameters such as the sampling frequency of optical identification and the transmission rate for audio communication.

Finally, it should be noted that although e-Vision, with the capabilities offered by CV, is oriented towards addressing practical needs of the visually impaired, as explained above, there is another, equally important benefit that the system provides to the visually impaired. While visually impaired people can often rely on the help of family members, friends or members of staff to accomplish the activities covered by e-Vision or other CV solutions, their deeper needs concern much more than the accomplishment of a task.

The autonomy of the visually impaired, expressed by their ability to accomplish everyday tasks on their own even when help by others is available, is extremely important. The ultimate benefit which e-Vision aspires to bring to the visually impaired is greater autonomy, and the increased wellbeing, which can be accomplished through it. To this end and in an effort to examine the actual benefits that e-Vision brings to the visually impaired community, a pilot-study was performed where ten visually impaired participants tried the introduced system in the three described scenarios (i.e. shopping at a grocery store, outdoor walking and visiting a public administration building). At the end of this, day-long study, the participants evaluated the system by means of answering corresponding questionnaires. All participants stated that e-Vision has the potential to benefit their everyday lives greatly once it reaches its release-ready version.

Panagiotis Migkotzidis holds a master's degree in Game Artificial Intelligence from the University of Malta. His main research interests are Game AI, Procedural Content Generation and AI assisted design tools. He is currently working as a research assistant in Information Technologies Institute (ITI) of the Centre for Research & Technology Hellas (CERTH), supporting the development of automated design tools.

Fotis P. Kalaganis is a Ph.D. student at Aristotle University of Thessaloniki, Department of Informatics. Meanwhile he is working as a research associate in Information Technologies Institute (ITI) of the Centre for Research & Technology Hellas (CERTH), developing signal processing algorithms and Brain-Computer Interfaces.

Kostas Georgiadis is a Ph.D. student at Aristotle University of Thessaloniki, Department of Informatics. Meanwhile, he is working as a research associate in Information Technologies Institute (ITI) of the Centre for Research & Technology Hellas (CERTH), developing signal processing algorithms and Brain-Computer Interfaces.

Elisavet Chatzilari received her diploma degree in Electronics and Computer Engineering from the Aristotle university of Thessaloniki (2008) and her PhD degree on social based scalable concept detection from University of Surrey in 2014. She is currently a post-doctoral research fellow at Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH).

George Pehlivanides holds an undergraduate degree in graphic information design and a postgraduate degree in communication design - pathway in interactive multimedia. He works as research associate and interaction designer in various research projects for the cultural sector.

Spyros Tsafaras attended the School of History and Archaeology of Aristotle University, where he also completed his postgraduate studies on Classical Archaeology. He has worked as an archaeologist and he has also collaborated in several research projects, all concerning the digitalization, documentation and promotion of the Greek cultural heritage and modern culture. Since 2016, he is member of TETRAGON SA, Greece, and responsible for R&D projects and programs.

Kostas Monastiridis is an experienced Unity Developer, working in the architecture & creative industry as well as in the Game Industry for the last 5 years. Skilled in Unity3D, with a strong focus on UI programming, systems integration, VR/AR/MR development. He holds a Master's Degree focused in Media Technology (Medialogy) - Games Specialization from Aalborg Universitet.

George Martinidis holds an undergraduate and postgraduate degree in psychology, a postgraduate degree in economics and politics, and a PhD on regional development. He has worked for the Major Development Agency of Thessaloniki (MDAT) as an external expert.

Spiros Nikolopoulos holds a PhD degree on Semantic multimedia analysis using knowledge and context, Queen Mary University of London (2012). He is currently a senior researcher in Information Technologies Institute (ITI) at the Centre for Research & Technology Hellas (CERTH).

Ioannis Kompatsiaris is a Research Director at CERTH-ITI and the Head of Multimedia Knowledge and Social Media Analytics Laboratory. His research interests include multimedia, big data and social media analytics, semantics, human computer interfaces (AR and BCI), eHealth, security and culture applications.

References

Alonso, Fernando, José L. Fuertes, Ángel L. González, and Loïc Martínez. “User-Interface Modelling for Blind Users.” In *Computers Helping People with Special Needs*, edited by Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer, 5105:789–96. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. https://doi.org/10.1007/978-3-540-70540-6_117.

Auvray, Malika, Sylvain Hanneton, and J Kevin O’Regan. “Learning to Perceive with a Visuo — Auditory Substitution System: Localisation and Object Recognition with ‘The Voice.’” *Perception* 36, no. 3 (March 2007): 416–30. https://doi.org/10.1068/p5631.

Bartz, Christian, Haojin Yang, and Christoph Meinel. “STN-OCR: A Single Neural Network for Text Detection and Text Recognition.” *ArXiv*, 2017.

Bourbakis, Nikolaos, Robert Keefer, Dimitrios Dakopoulos, and Anna Esposito. “A Multimodal Interaction Scheme between a Blind User and the Tyflos Assistive Prototype.” In *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, 487–94. Dayton, OH, USA: IEEE, 2008. https://doi.org/10.1109/ICTAI.2008.52.

Brady, Erin, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. “Visual Challenges in the Everyday Lives of Blind People.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2117–26. Paris France: ACM, 2013. https://doi.org/10.1145/2470654.2481291.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “ImageNet: A Large-Scale Hierarchical Image Database.” In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–55. Miami, FL: IEEE, 2009. https://doi.org/10.1109/CVPR.2009.5206848.

Georgiadis, Kostas, Fotis Kalaganis, Panagiotis Migkotzidis, Elisavet Chatzilari, Spiros Nikolopoulos, and Ioannis Kompatsiaris. “A Computer Vision System Supporting Blind People - The Supermarket Case.” In *Computer Vision Systems*, edited by Dimitrios Tzovaras, Dimitrios Giakoumis, Markus Vincze, and Antonis Argyros, 11754:305–15. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-34995-0_28.

Gold, D., and H. Simson. “Identifying the Needs of People in Canada Who Are Blind or Visually Impaired: Preliminary Results of a Nation-Wide Study.” *International Congress Series* 1282 (September 2005): 139–42. https://doi.org/10.1016/j.ics.2005.05.055.

Jourabloo, Amin, and Xiaoming Liu. “Pose-Invariant Face Alignment via CNN-Based Dense 3D Model Fitting.” *International Journal of Computer Vision* 124, no. 2 (September 2017): 187–203. https://doi.org/10.1007/s11263-017-1012-z.

Kuznetsova, Alina, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, et al. “The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale.” *International Journal of Computer Vision* 128, no. 7 (July 2020): 1956–81. https://doi.org/10.1007/s11263-020-01316-z.

Martin, Robert C. “Design Principles and Design Patterns.” *Object Mento*r 1, no. 34 (2000).

Mirandola, Daniela, Marco Monaci, Guido Miccinesi, Alessia Vannuzzi, Eleonora Sgambati, Mirko Manetti, and Mirca Marini. “Psychological Well-Being and Quality of Life in Visually Impaired Baseball Players: An Italian National Survey.” Edited by Stefano Federici. *PLoS One* 14, no. 6 (2019): e0218124. https://doi.org/10.1371/journal.pone.0218124.

Olofsson, Stina. “Designing Interfaces for the Visually Impaired : Contextual Information and Analysis of User Needs.” Umeå University, 2018.

Qin, Zheng, Zhaoning Zhang, Xiaotao Chen, and Yuxing Peng. “FD-MobileNet: Improved MobileNet with a Fast Downsampling Strategy.” In *2018 25th IEEE International Conference on Image Processing* (ICIP), 1363–67. IEEE, 2018.

Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. “You Only Look Once: Unified, Real-Time Object Detection.” In *2016 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 779–88. Las Vegas, NV, USA: IEEE, 2016. https://doi.org/10.1109/CVPR.2016.91.

Sierra, Javier Sánchez, and Joaquín Selva Roca de Togores. “Designing Mobile Apps for Visually Impaired and Blind Users: Using Touch Screen Based Mobile Devices: IPhone/IPad.” *ACHI 2012: The Fifth International Conference on Advances in Computer-Human Interactions*, 2012, 47–52.

Sun, Yi, Xiaogang Wang, and Xiaoou Tang. “Deep Convolutional Network Cascade for Facial Point Detection.” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, 3476–83.

Walker, Jake, Yasuhisa Fujii, and Ashok C. Popat. “A Web-Based OCR Service for Documents.” In *Proceedings of the 13th IAPR International Workshop on Document Analysis Systems* (DAS), Vol. 1. Vienna, Austria, 2018.