Computer Vision and Image Understanding 149 (2016) 209–228

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

An integrated artificial vision framework for assisting visually impaired users

Manuela Chessa[1,*], Nicoletta Noceti[1], Francesca Odone, Fabio Solari, Joan Sosa-García, Luca Zini

DIBRIS - Università degli Studi di Genova via Dodecaneso, 35, 16146 Genova, Italy

ARTICLE INFO

Article history:
Received 17 April 2015
Accepted 18 November 2015

Keywords:
Applications for the visually impaired
Scene understanding
Surface orientation
Time to collision
Face detection and recognition
Text detection and recognition
Object recognition
Bio-inspired computer vision

ABSTRACT

We present a conceptual framework inspired by biological vision which integrates low-level vision functionalities oriented to actions – typical of the so-called "where" dorsal pathway – with identification and recognition capabilities – common to the "what" ventral pathway. Although they proceed independently, these complementary vision models may provide a deeper scene understanding and a more efficient computational framework. In this work, we refer specifically to a set of video analysis modules, which include semantic annotations of the scene and 3D environment interpretation. We discuss and qualitatively evaluate possible connections and integrations between different functionalities, grounding our analysis on a set of specific use cases, depicting visually impaired users finding their way in unfamiliar environments.

1. Introduction

Visually impaired people find it difficult to operate in unfamiliar environments or unpredictable settings. These challenges limit their independence and widen the gap between them and the normally sighted population. They are likely to modify their life to adapt to this particular condition, limiting their actions to prevent failures and frustrations. Usually, after years of training they learn how to conduct a reasonably independent life, in known environments and under a repetitive schedule. Often, they need to be taken to places for many times until they learn by heart all the important details of a given path, using complementary senses in the case of severe visual deficiencies [1]. In less controlled circumstances, they may have problems in finding their way in unknown places or in avoiding physical barriers. Moreover, it may be difficult for them to choose clothes and to buy objects; it is challenging to find a shop of a specific type in a street they do not know well, or to enter a post office, as they do not know if the clerk is ready to serve them.

According to the European Blind Union[2], there are nearly 30 million blind and partially sighted people in the member countries of the European Blind Union. This figure is based on the premise that 1 in 30 people are blind or partially sighted, and takes into account the varying definitions of visual impairment. This number is expected to grow in the future, together with the increase of the citizens' average age. At the same time, we can assume the number of elderly people willing to use technology will also grow.

Therefore, in the last few years there has been a steady growth of research and development of methods and systems to assist visually impaired people. While there has been a considerable advancement at the level of research findings, still most issues have been addressed independently. This lead to the development of devices or dedicated apps that may, in principle, be considered interesting aids, but it is virtually impossible to imagine a user adopting them in parallel. Only recently, some products are starting to propose multiple functionalities (e.g. the ORCAM device, which appears to be largely based on an effective OCR).

In this paper, we explore the possibility of designing multifunctional aids, purely based on Computer Vision, to address general goals of scene understanding and wayfinding. We reason on the benefits of integrating different functionalities in the same framework, obtaining improved performances, a wider descriptive power, and a potential for developing more general purpose devices in the future.

In our work, we draw inspiration from biological visual processing. We refer to the neural modeling of the vision functionalities, that states how any living being that acts in the environment has the need of visually exploring the neighboring locations, in order to both navigating and recognizing objects. Such an ability has evolved in two

* Corresponding author.
E-mail address: manuela.chessa@unige.it (M. Chessa).

[1] Manuela Chessa and Nicoletta Noceti equally contributed to the paper.

[2] The European Blind Union (EBU) is a non-governmental and non-profit-making organisation founded in 1984, currently composed of 45 countries: http://www.euroblind.org/.

http://dx.doi.org/10.1016/j.cviu.2015.11.007
1077-3142/

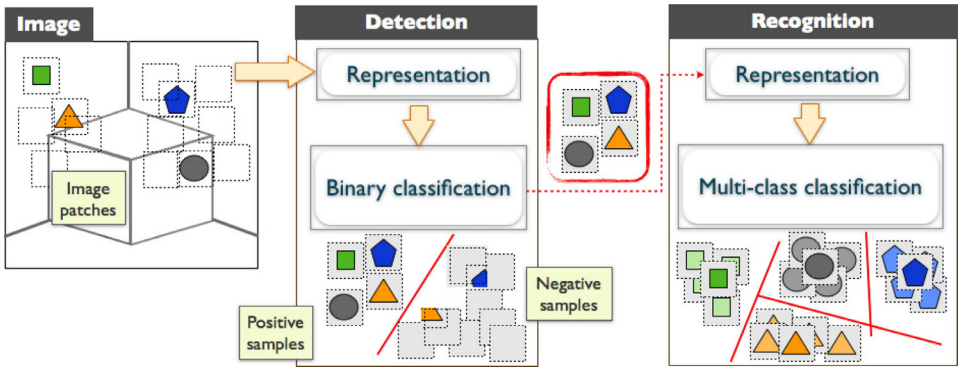

Fig. 1. A visual representation of detection and recognition pipeline. The actual detection relies on classifying image patches using a binary classifier that decides whether the patch contains or not an object of interest (e.g. shapes). Only positive samples undergo the recognition step, based on a multi-class classification, to associate a sample with an object class (e.g. square, circle,...).

different visual pathways in the primate cerebral cortex [2]. The ventral visual pathway (the *"What" stream*) [3] is devoted to build a detailed representation of the environment features required for cognitive operations, such as identification and recognition. The dorsal visual pathway (the *"Where" stream*) [4] provides a distributed representation of visual features, such as objects motion and distance, which enables the visual control of actions. In particular, such visual pathways are organized in hierarchical layers of neural architectures performing visual processing that becomes more informative and complex in higher layers.

The framework we propose has the aim of mimicking the first visual processing stages, which are present in the dorsal and ventral streams, in order to provide early visual capabilities to impaired people. Figs. 1 and 7 show the proposed instances of the "What" and "Where" visual pathways, respectively.

Although we do not consider a specific hardware, our reference technologies are wearable devices, such as sensorized glasses (usually mounting one camera). We simulate this configuration by processing video streams acquired by mobile phones.

We consider a few reference functionalities, which are summarized in the following. They are not intended to be an exhaustive list, but rather a rich starting point for future investigations.

Analysis of the environment. The physical constraints of an environment are 3D cues that normally sighted people perceive and understand very easily. For visually impaired people, instead, the same perceptual task becomes highly challenging, especially in unfamiliar places. In this respect, we provide an intuitive and simple way to obtain coarse information about the 3D structure of the scene (e.g. noticing the presence of walls and obstacles, and their temporal distances) in the user's field of view. The analysis of the environment is performed by a two-stage neural model, which first estimates low-level visual features, related to motion and depth, and then builds a sparse representation of the object surfaces and of the time to contact. Such information allows visually impaired people to be aware of the spatio-temporal structure of the scene, and, thus, to move in unknown environments.

Text reading. Natural scenes contain large amount of information directed to humans which is conveyed through text. The capability of "reading" textual information may favor visually impaired users moving in the environment for locating a specific shop, or allowing them to localize themselves by reading the streets names. We provide methods relying on a two-stage solution that first detects and then recognizes characters. Each character is processed independently. A reading functionality may be applied, in order to finally recognize words.

Object recognition. In everyday life, we often need to recognize and use large varieties of objects. In familiar settings, a visually impaired user may learn the position of specific objects by heart. When interacting with the external world, the same problem may be very challenging. Their limited visual ability triggers the refinement of alternative perception capabilities (e.g. tactile), and in some cases exploring the shape of a 3D object is sufficient to achieve the recognition. However, there are circumstances where the only shape is not informative enough. In such cases, the user may be able to roughly categorize the object (e.g. deciding that it is a box or a can) but cannot understand more specific properties (e.g. the content). As a use case for this wide scenario, we consider the problem of banknotes recognition. The same pipeline may be used to address a larger class of recognition problems, which aim at identifying objects observed by a close range.

Face detection and recognition. An important and quite general difficulty for blind users is in interacting with known and unknown people in unfamiliar environments. In the former case, they do not have clues on the presence of a known person in the environment, unless the other person sees them and sends them a vocal feedback; in the latter case, they may not know there is a person in their vicinity or they do not know whether this person is paying attention to them or not. This lack of visual contact greatly reduces the spontaneity of interpersonal contacts. We then consider a fast and effective face detection and recognition pipeline. The face detection module alerts the user on the presence of a person in the surroundings; in conjunction with time to contact, it may alert on the fact that the person is approaching. The face recognition module informs the user if a known person is present in the scene. It is possible to "introduce" a new person to the system for future occasions.

The reminder of the paper is organized as follows. In Section 2 we briefly review the literature of related works addressing assistive vision problems. Section 3 provides an overview of our methods for assigning semantic tags to a scene, with specific reference to two different scenarios of interest; Sections 4 and 5 are focused on the introduction of methods we used to extract low-level features, and on how we exploit them for understanding the coarse structure of the scene, respectively. Section 6 provides examples of integrated functionalities which are shown on a selection of use cases, while Section 7 is left to conclusion.

2. Related works

In the last decades, much attention has been devoted to providing technological aids to visually impaired users.

Many navigation devices have been devised to guide blind people, some of them based on non-visual sensors, such as laser emitters and