Sainet: An Image Processing App for Assistance of Visually Impaired People in Social Interaction Scenarios

Jesus Salido, Oscar Deniz, and Gloria Bueno(✉)

Department of IEEAC, Castilla-La Mancha University, Ciudad Real, Spain
{jesus.salido,oscar.deniz,gloria.bueno}@uclm.es
http://visilab.etsii.uclm.es/

Abstract. This work describes a mobile application (Sainet) for image processing as an assistive technology devoted to visually impaired users. The app is targeted to the Android platform and usually executed in a mobile device equipped with a back camera for image acquisition. Moreover, a wireless bluetooth headphone provides the audio feedback to the user. Sainet has been conceived as an assistance tool to the user in a social interaction scenario. It is capable of providing audible information about the number and position (distance and orientation) of the interlocutors in the user frontal scenario. For validation purposes the app has been tested by a blind user who has provided valuable insights about its strengths and weaknesses.

Keywords: Image processing · Mobile computing · Visually impaired assistance

1 Introduction

In the last few years mobile devices have become a pervasive technology in our world. Together with the increasing use of mobile devices the images taken by them are reaching one trillion during 2015. Nowadays the pictures and videos captured by smartphones and tablets surpass those taken by digital cameras while the number of mobile apps available in the online stores is almost 4 millions (www.statista.com). In such scenario apps for image and video processing are reaching a significant role.

Usually the computer vision software for real time processing requires an expensive and high performance hardware. Today the hardware required for astonishing mobile apps capable of video and image real time processing is available in our hands thanks to multi core processing units and high resolution cameras present in the current commercial smartphones.

In addition to hardware, software also facilitates development of computer vision software on mobile devices. Since OpenCV [1,2,4] was released, this open source library has become the standard in the software development community and hopefully available for the main mobile platforms. Besides being a toolbox

F. Ortuño and I. Rojas (Eds.): IWBBIO 2016, LNBI 9656, pp. 467–477, 2016.
DOI: 10.1007/978-3-319-31744-1_42

for computer vision, OpenCV constitutes an standard benchmark to compare different solutions to the same problem.

In this work the VISILAB Group[1] has been exploring the capabilities of mobile computer vision as an assistive technology for the visually impaired (VI). The technologies for assistance of VI people have been around for decades [5–7,12] and the solutions are classified into three categories:

1. Vision sense enhancement. They are devoted to image capture to process it and to display it in such a way that be more easily perceived (as augmented or sharpened on screen).
2. Vision sense recovery. They are systems devoted to driving visual signals directly to the neural cortex.
3. Vision sense substitution. Similar to the first one, however in this case visual information is converted to auditory or tactile information.

Multiple integrated solutions have been provided that might be considered into the category of Vision sense substitution:

– Electronics Travel Aids (ETAs). They are systems to support navigation, usually as substitutive of the white can. The information about the environment can be gathered by cameras, laser scanners or sonars.
– Electronic Orientation Aids (EOAs). They are devices that provide orientation prior to, or during the movement. They can be external to the VI user and/or can be carried by the VI user (e.g., infrared light transmitters and handheld receivers).
– Position Locator Devices (PLDs). When used outdoors they rely on technologies like GPS. Localization in indoor environments still remains a challenge.

In all the mentioned systems the information supplied by human vision is substituted by hearing and touch inputs. There is a general agreement about the requisites that solutions should meet depending on the viewpoint. From the VI user perspective: fast feedback to the user allowing real time operation without interfering other sensory inputs, unobtrusive and lightweight so that they can be carried during long distances and times, reliable even under unexpected circumstances, affordable for most users, easy to use and having valuable functionality. From the developer side the requirement should be: simplicity, robustness, connectivity, performance, originality and improvement capability.

The previous classification on three categories (ETAs, EOAs and PLDs) for vision substitutive systems does not include systems aiming to facilitate daily tasks (e.g. buying in a marketplace) where object and people recongnition is desirable for goal achievement and fulfil the social interactions required [3,14–18]. In particular this work addresses the assistance to blind people in social interaction scenarios which had been subject for a short extent of previous efforts [9,10].

In social sitiuations VI people do not have visual cues and they would value information about their interlocutors such as: appearance, position, facial expressions, and so on. For example a typical situation is the inference of the distance to

[1] Vision and Intelligent Systems Research Group at University of Castilla-La Mancha.

our interlocutor and his/her degree of interest on the conversation which might be a cue to change or stop the conversation. At times, the scarceness of such information may derive on social exclusion and alienation [9,10] assuming that at least 65 % of a two-person conversation is non-verbal.

The majority of works focused, until now, to VI users consist on very specific "portable" devices. Only recently the smartphone has been considered as a feasible assistive technology [5–7,12]. In fact, it has become not only an outstanding contender for mobile assistive technologies but one of the principal *mHealth* platforms [13]. Specially for VI assistance, the smartphone is having a main role derived from its ability to embed heavy processing tasks as required by computer vision applications.

2 Objective

This work describes the development of Sainet, a mobile application for image processing with assistive purposes for VI people in social interaction scenarios [9,10,12]. The app includes vision algorithms to allow people detection (number of persons) and estimation of distance and position of individuals in front of the VI user. The scene in front of the VI user is captured by the back camera of the VI user's smartphone while the information is converted to audible feedback supplied to the VI user by a portable earphone (wirelessly or not).

The main innovative aspects of the proposed solution with Sainet are:

1. Portability. The application is embedded into a commercial smartphone with VI user feedback by an earphone. In that way the solution becomes unobtrusive and easy to use.
2. Devoted to provide visual information related to the social interactions. As social interactions deal with people on the surrounding environment, people detection is one of the key points.
3. Without interference to other sensory inputs. The visual information must be fed back to the VI user as a sound input, however it must be done carefully to cope with possible overload of the auditory channel.
4. Real time operation. The system developed should drive feedback at almost the same pace as social interactions take place, although some task are more demanding than others (i.e. face detection versus gesture detection). For a computer vision app this means being able to process several frames per second (fps).
5. Scalability. Although the objective is very concrete, the app has been conceived as a platform to easily incorporate new funcionality (e.g. smile detection, gesture detection, face recognition, etc.).

3 Requirements and Dependencies

The app requirements and dependencies include the operating system (OS) and auxiliary software that have to be installed in the VI user smartphone:

– OS Android 4.0+. This choice relies on the principal adoption of this OS in the Spanish market (about 92 %). OS version dependency is important because the native API (Application programming Interface) is used for face detection.
– Auxiliary software.
  • Text-to-Speeech (TTS) synthesis engine (for instance PicoTTS). It is responsible of the auditory feedback to the VI user through voice messages. This engine is included with the OS and it can be also downloaded from Google Play.
  • OpenCV Manager [1,2,4]. It is the OpenCV binary core library for Android downloaded from the Google App Store the first time the app Sainet is executed.

Fig. 1. The VI user testing the Sainet app

The hardware requirements for the mobile phone are truly basic and fulfilled by a wide range of available devices at the present market:

– Back camera. It captures the scene in front of the VI user.
– Bluetooth. It allows wireless connectivity between the smartphone and a wireless earphone for auditory messaging feedback to the user.
– Earphone. It is recommended a wireless earphone for a more unobtrusive set up, however a wire plug earphone is also possible.

In the validation test for Sainet (see Fig. 1) a mobile phone with the following characteristics has been used: model HTC One S (Z520e) 1.5 GHz dual core with 1 GB of RAM and a rear camera with 8 Mega pixels (f2.0, 28 mm, AF).

4 Functional Description

Sainet has been developed with the Java language programming using the Android native APIs for TTS and face detection on the images captured. Moreover the OpenCV library is used for torso detection allowing a more robust detection of people on the scene.

Once the application is initiated, a continuous loop with the following four stages is executed:

1. image capture,
2. person localization by face detection,
3. position and distance estimation, and
4. auditory feedback if the user requests it.

4.1 Image Capture

The image is captured by the smartphone back camera. As soon as one frame is processed another is captured trying to get the highest rate. Each camera provides different frame resolutions (i.e. number of pixels) that affect the amount of memory needed and the processing time. By default the frame resolution is chosen at mid-range, however this value can be modified (e.g. in the tested device resolutions can be selected among 1280×720, 480×20 and 176×144 pixels).

4.2 Person Localization by Face Detection

Person localization is a key problem in social interaction. This is a common problem in other computer vision scenarios (e.g. surveillance). The solution to the problem relies on detecting each person on the image and tracking them while they are in the scene. Many works have been devoted to cope with this problems [9]. The simple approach adopted in the Sainet system derives from two assumptions:

1. Close face-to-face social interactions are the most important. In this context 'close' means 3 meters or less.
2. The VI user needs assistance only when visual information is relevant and unreachable by other sensory inputs.

Human body detection on images is a tough task because the non rigid shape of human body adds a new challenge to the object detection problem. However under the first assumption listed above, person detection can be solved by face detection. Then *'person localization by face detection'* becomes an easier problem with available implementations in both the Android native and OpenCV APIs.

Face detection with OpenCV is achieved using Viola-Jones algorithm [8] that uses Haar cascade of classifiers. This algorithm applies naive classifiers to detect single features (i.e. Haar-like features) on the image building an 'stronger' classifier for face detection by training an AdaBoost system. The Viola-Jones algorithm has a very high rate of accuracy, offering about 1 % or less for false negatives (i.e. missing a face that is present) and under 40 % for false positives (detecting a face when there is none).

The main advantages of Viola-Jones algorithm are:

– Availability of open source implementation (e.g. in OpenCV) being freely to use.

– Short processing time. Around 3 fps (aprox. 300 ms between consecutive captures) in the smartphone used for testing. This time depends on frame resolution and the performance of the processing unit.
– Far from 'close' detection. Depending on frame resolution the algorithm can detect faces up to 8 m for a 1920 × 1080 frame resolution under testing conditions.

Android also provides a native algorithm for face detection. The algorithm was patented by Neven Vision, a company acquired by Google in 2006. This algorithm can be only used as part of the Android API and his performance (5 fps) surpass the Viola-Jones implementation in OpenCV for Android in the testing device. The Neven algorithm offers a quite independent accuracy from frame resolution selected under test, however its detection range degrades beyond 3 m.

The main drawback observed for the two aforementioned algorithms arises from the moderately high false positive detection rates. In testing conditions false positives produce too frequent 'phantoms' that generate erroneous useless feedback to the VI user. To decrease the false positive rate (i.e. increasing detection robustness) the Sainet system adds a confirmation stage based on torso detection implemented with a Haar filter cascade implemented with the OpenCV library. The combination of face and torso detection offers a more robust detection with a total processing time around 2 fps.

4.3 Position and Distance Estimation

Once the persons are detected on the image the step is to estimate their distance and position relative to the observer. An accurate estimation is difficult to reach from a monocular sequence of frames. However, metric accuracy on distance and position estimation is meaningless from the viewpoint of a VI person. From the VI user perspective a rough estimation is in fact more significant. Furthermore this kind of estimation is quite fast to obtain.

Because the social interaction of interest covers a range of 3 m from the VI user, distance and position may be quantized onto three values respectively (*'next'*, *'close'* and *'away'* for distance, and *'left'*, *'central'* and *'right'* for position). Thereby the scene is divided in a simple 3 × 3 shaped space. To estimate position the image is evenly divided in three vertical regions (see Fig. 2) while distance measure is obtained by setting two thresholds for the area of face detection (i.e. more area for closer face detection).

4.4 Auditory Feedback to the VI User

When the required information is extracted from the image in the previous step, it is time to feedback this information to the VI user by a different communication channel rather than visual. The alternative interfaces reported in the specialized literature [5–7,11,12,16] can be classified in two categories: tactile (or haptic) interfaces and sound interfaces. None of these categories is optimal because they overlap the substitute senses of vision for VI users.

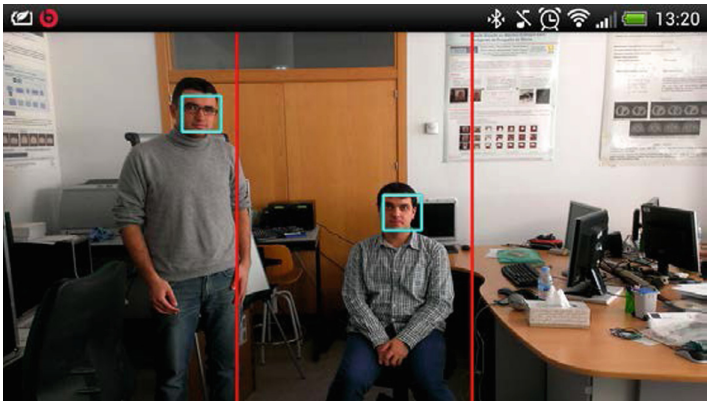

Fig. 2. Detection visualization and discrete scales of position a distance (just for testing purposes)

In Sainet, speech synthesis was chosen as interface with the VI user. Rather than a continuous feedback to the user, the information is only sent when the user requests it. Whenever a request is initiated by the user, the system converts the information gathered from the scene into a voice message. Wearing an earpiece properly connected to the smartphone, the user listens the voice message played by the system (e.g. *"Next to two people at the right. One more to the far left"*).[2]

The main design decisions around this interface can be explained as follows:

- Sound vs. Haptic. Depending on the familiarity of sound this kind of interface requires short time for training (that may be untrue for an odd tone codification). Moreover sounds interfaces are easier to implement because they do not require an specific device rather than the smartphone and an earpiece.
- Voice vs. tone codification. During previous stages of the work a sound codification, rather than synthetic voice, was tested. In this interface a sound localization technique provides the spatial awareness about distance and position. This solution was rejected during validation with the VI user because the complexity for system calibration and the saturation obtained for the auditory sense. On the contrary, synthetic voice is inherently easy to understand without training nor calibration stages. In the final interface a module generates a text message sent to the Android TTS (Text-to-Speech) engine. The inputs to the module are the number of individuals in scene, their distance and position.
- Feedback under request vs. continuous. One of the principal requisites for interfaces devoted to VI users is that they do not interfere with their remain senses, specially touch and hearing. A continuous artificial feedback to the user usually tends to saturate the communication channel and cause stress on VI users (as reported in validation stages with Sainet prototype). Then, a more natural approach results from an strategy of *'feedback under request'* (i.e. *"get information only when needed"*).

[2] This message is a translation for the actual Spanish implementation.

Fig. 3. User validation of detection under request

In the *'feedback under request'* approach the unsolved question is: *how to notify the request to the system?* Several options were tested with the VI user: gestures on terminal, pressing of smartphone's volume buttons (see Fig. 3), use of NFC technology (Near Field Communication) and miniature bluetooth switches. The next section explains the validation process and discuss the main lessons learned.

5 Validation and Discussion

From the beginning, Sainet was focused on the end-user. Therefore, a VI user (blind) was included as a team member for co-developing the SAINET project. His role was to participate in each validation step of the development moreover his experience becomes a source of endless insight to direct the team's efforts. The system validation was organized in two stages directed to gather the experience from the user. First stage was *'practice with the system'* and second, *'explaining the experience'*. Before the experience began, the user was informed about the purposes of the experience and questions are answered about how to *'play with'* the system. After the experience, the team had a meeting session driven by a questionnaire to the user for recording his answers, comments and suggestions.

The main aspects validated by the user were:

- Robustness of person detection. The strategy combining face and torso detection is helpful under not extreme illumination conditions even in outdoors scenarios, although detection rates are affected by illumination conditions and relative movement between camera and individuals.
- Auditory feedback of scene information. The mobile app (i.e. Sainet system) provides a valuable information about position and distance of people in the scene. Rather than continuous, this info is fed back to the VI user in a *'request*

under demand' mode so that the user can decide when to be informed by the system. After that, the information extracted from the scene was converted into a single voice message.
– Application accessibility. Although Sainet does not aim to provide additional accessibility mechanisms to the available ones from the Android OS (e.g. TalkBack, quick widget launcher, etc.), it is compatible with such as mechanisms.

Among all the lessons learned by the Sainet team with the help of our mate (i.e. the VI user), several statements can be established as principles of assistive technologies for the VI:

1. The substitutive innate senses of a VI user must be free as much as possible. The assistive technology should never saturate nor interfere with the natural senses of the user. Under this principle a *'request under demand'* mode was adopted in the Sainet system and a alternatives based on audio 3D and continuous feedback were discarded.
2. The assistive system must be easy to use. To promote the use of technology it has to be user friendly. In Sainet, voice messages generated by a TTS engine proved to be quite simple and intuitive as the way for communicating the relevant information to the user.
3. The assistive technology should be unobtrusive to interlocutors. To make VI users feel self-confident the assistive technology needs to be easy to wear and unobtrusive. Smartphone use is common even in social interaction scenarios, however Sainet only meets this principle partially. In the user opinion the system should be improved in two related aspects:
   (a) Image capture. The image capture in Sainet is done with the smartphone located on the user torso, to use the back camera as capture device. However, this configuration raises some objections derived from this unnatural setting. In fact, during validation the VI user decided by himself just to take the smartphone in his hands holding it at his eyes height (see Fig. 3). The favourite alternative from the user's perspective would be a capture device included in the protective glasses they usually wear. This option requires a miniature camera with wireless connection to the smartphone where the image would be processed. In this solution the glasses could also include the earphone.
   (b) Feedback request. Several options were tested to provide a request mechanism in Sainet system: pressing of volume buttons, terminal gestures, triggering with NFC tags, and bluetooth clickers. Although the bluetooth clickers provide an unobtrusive mechanism for triggering a request, the VI user felt unwilling to include additional devices to the system. In this sense an smartwatch could play the role of triggering requester without adding auxiliary devices for this exclusive purpose.

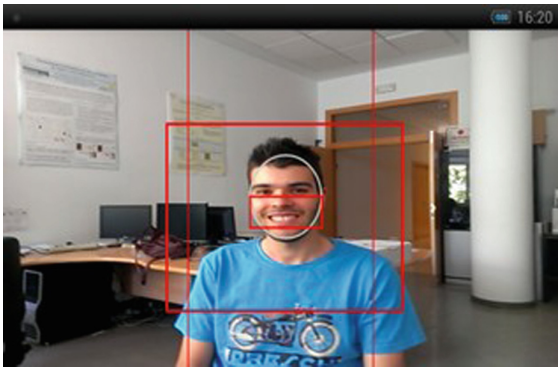

Fig. 4. Smile detection in the image captured

6 Conclusion

After VI user validation it was concluded that Sainet brings an easy usage without demanding special requirements from present mid-range marketed smartphones. The application provides valuable information about the scenario captured by a mobile camera, being also scalable to incorporate new functionality. The additional functionalities considered for futures versions are:

1. Recognition of facial expressions, hands and head gestures (including gaze direction). To investigate the feasibility of new feature addition in the system an experimental smile detection feature was included in the final version (see Fig. 4).
2. Automatic text detection and reading (i.e. posters, bills, product labels, etc.). These functionalities are very common in daily tasks with valuables outcomes for the VI users.
3. Re-recognition of previously tagged people and objects.

Acknowledgments. This work describes the results for the project SAINET funded by a grant from the Indra-UCLM university Chair and the Adecco Foundation. The authors want to acknowledge the received collaboration from the VISILAB Research Group and specially to Sergio Vera, Francisco Torres and Jesús Manzano.

References

1. Baggio, D.L., Emami, S., et al.: Mastering OpenCV with Practical Computer Vision Projects. Packt Publishing, Birmingham (2012)
2. Deniz, O., Salido, J., Bueno, G.: Programación de Apps de Visión Artificial. Bubok Publishing S.L. (2013). http://visilab.etsii.uclm.es
3. Deniz, O., et al.: A vision-based localization algorithm for an indoor navigation app. In: 8th International Conference on Next Generation Mobile Applications, Services and Technologies, pp. 7–12. IEEE (2014)

4. Bueno, G., Deniz, O., et al.: Learning Image Processing with OpenCV. Packt Publishing, Birmingham (2015)
5. Manduchi, R., Coughlan, J.: (Computer) vision without sight. Commun. ACM **55**, 96–104 (2012)
6. Dakopoulos, D., Bourbakis, N.G.: Wearable obstacle avoidance electronic travel aids for blind: a survey. Trans. Syst. Man Cybern. Part C **40**(1), 25–35 (2010)
7. Velázquez, R.: Wearable assistive devices for the blind. In: Lay-Ekuakille, A., Mukhopadhyay, S.C. (eds.) Wearable and Autonomous Biomedical Devices and Systems for Smart Environment. LNEE, vol. 75, pp. 331–349. Springer, Heidelberg (2010)
8. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. (IJCV) **57**(2), 137–154 (2004)
9. Gade, L., Krishna, S., Panchanathan, S.: Person localization in a wearable camera platform towards assistive technology for social interactions. Special Issue on Media Solutions that Improving Accessibility to Disabled Users, Ubiquitous Computing and Communication Journal (2010)
10. Krishna, S., Colbry, D., et al.: A systematic requirements analysis and development of an assistive device to enhance the social interaction of people who are blind or visually impaired. In: Workshop on Computer Vision Applications for the Visually Impaired Conducted Along with European Computer Vision Conference (ECCV), Marseille, France (2008)
11. Krishna, S., Panchanathan, S.: Assistive technologies as effective mediators in interpersonal social interactions for persons with visual disability. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) ICCHP 2010, Part II. LNCS, vol. 6180, pp. 316–323. Springer, Heidelberg (2010)
12. Terven, J.R., Salas, J., Raducanu, B.: New opportunities for computer vision-based assistive technology systems for the visually impaired. J. Comput. **4**, 52–58 (2014). IEEE Computer Society
13. Becker, S., Miron-Shatz, T., et al.: mHealth 2.0: experiences, possibilities, and perspectives. JMIR mHealth uHealth **2**(2), 1–12 (2014)
14. Zhu, X., Ramanan, D.: Face detection, pose estimation and landmark localization in the wild. In: Computer Vision and Pattern Recognition (CVPR) (2012)
15. Yi, C., et al.: Finding objects for assisting blind people. Netw. Model. Anal. Health. Inform. Bioinf. **2**, 71–79 (2013). Springer
16. Ivanchenko, V., Coughlan, J.M., Shen, H.: Crosswatch: a camera phone system for orienting visually impaired pedestrians at traffic intersections. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) ICCHP 2008. LNCS, vol. 5105, pp. 1122–1128. Springer, Heidelberg (2008)
17. Yang, X., Tian, Y.: Robust door detection in unfamiliar environments by combining edge and corner features. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 57–64 (2010)
18. Winlock, T., Christiansen, E., Belongie, S.: Toward real-time Grocery detection for the visually impaired. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 49–56 (2010)