Usage of Artificial Vision Cloud Services as Building Blocks for Blind People Assistive Systems

Dennis Paulino, Arsénio Reis, Hugo Paredes, Hugo Fernandes, João Barroso

Abstract: ***This study has the objective of select the best service at image processing and recognition, running in the cloud, and best suited for usage in systems to aid and improve the daily lives of blind people. To accomplish this purpose, a set of candidate services was built, including Microsoft Cognitive Services and Google Cloud Vision. A test mobile app was developed to automatically take pictures, which are sent to the online cloud services for processing. The results and the functionalities were evaluated with the aim to measure their accuracy and relevance. The following variables were registered: relative accuracy, represented by the ratio of the number of accurate results vs. the number of results shown; confidence degree, representing the service accuracy (when provided by the service); and relevance, identifying situations that can be useful in the daily lives of the blind people. The results have shown that these two services, Microsoft Cognitive Services and Google Cloud Vision, provided good accuracy and significance, in supporting systems to help blind people in their daily tasks. It was chosen some functionalities in two APIs of services running in the cloud like face identification, image description, objects, and text recognition.***

Keywords: Blind people Cloud services Image recognition Mobile apps Android

I. INTRODUCTION

Presently, there are approximately 40 million people blind in the world. The loss of vision causes significant human suffering to the affected individuals and their relatives, conditioning their autonomy as well [1], so there is an increasing interest in technologies to assist blind people, providing more autonomy and improving their daily lives.

The recent advances in portable devices technology, together with advances in algorithms and systems dedicated to computer vision, is creating the basis for new applications of computer vision as an assistive technology for blind people [2-3]. This approach is being used to create interactive and intelligent software assistants, using cloud based cognitive services and specially target to address specific issues [4]-[5].The advances in technologies have brought more coherence in several tasks such has recognizing text and objects. A key advance on mobile vision technology are the data processing services in the cloud, allowing thinner devices, such as smartphones, to collect images and videos and have them processed on the cloud, using powerful algorithms and benefiting from big data analysis, machine learning, deep learning and several other developments related with weak artificial intelligence (AI).

Big data consists on large volumes of complex data analyzed using machine learning and deep learning algorithms as well as their implementation on consumer grade computers [6].

Machine learning algorithms are being widely adopted in computer science, as it allows systems to solve problems based on generalization from previous examples [7]. Within the machine learning research community, there is a specific trend of deep learning that advocates the design of algorithms with multiple levels of representation capable of catching finer and more subtle details from example and generalization [8].

Some of the current advances in Artificial Intelligence (AI) are in image processing, mainly focused on perception and interpretation of images. These are only a few aspects of intelligence and are considered as weak AI, while a strong AI would be a self-conscientious system, capable of thinking the same way as humans [9]-[10]. These weak IA technologies of image processing, are now implemented as data processing services in the cloud. The purpose of this work is to assess the usage of the currently available services as components of systems to help blind people.

To evaluate these services, it was set the main goal of providing autonomy to bling people in their daily lives and it was created a test environment, comprising several appliances and an application, that would interface with the user and would integrate the more useful and accurate features, provided by the currently available cloud services. The app automatically: 1. takes pictures; 2. sends them to the external cloud services for processing; 3. receives the results and sends then to a user device, which will communicate the results to the blind user as audio messages.

The user device runs a module of the CE4BLIND [[11]] project, which generates and manages the audio messages to the blind user. CE4BLIND is an ongoing project to create an assistive platform for blind people [12].

This paper is arranged in the following sections: “Background”, providing context to this work and comparing the currently available services running in the cloud that have some features related to this work; ”The testbed system”, describing the system that was developed to test the cloud services; “Methodology”, describing the method used to conduct the tests; “Tests and results”, reporting the results obtained and their analysis; ”Conclusions and discussion”, which summarizes the work and what was learned; and “Future work”, providing a guide for the following research.

Revised Manuscript Received on September 25, 2019

Dennis Paulino, INESC TEC and University of Trás-os-Montes e Alto Douro, Vila Real, Portugal
Arsénio Reis, INESC TEC and University of Trás-os-Montes e Alto Douro, Vila Real, Portugal
Hugo Paredes, INESC TEC and University of Trás-os-Montes e Alto Douro, Vila Real, Portugal
Hugo Fernandes, INESC TEC and University of Trás-os-Montes e Alto Douro, Vila Real, Portugal
João Barroso, INESC TEC and University of Trás-os-Montes e Alto Douro, Vila Real, Portugal

II. BACKGROUND

The project on which this article is based is part of a major project, the CE4BLIND Project that has the general objective of increasing the autonomy of blind people with technology aid. CEBLIND is a platform in development by INESC-TEC [13] and includes several components, such as: Blavigator [15], navigation system for blind people using an electronic cane; Smartvision [16], that uses computer vision techniques for object detection and obstacle detection; and an artificial vision module that we developed to test the cloud services reported in this article.

Currently, there are multiple image processing services that could be used in systems to help blind people. Some services running in the cloud are: Clarifai [17]; Microsoft Cognitive Services [18]; Google Cloud Vision [19]; and Abby Cloud Services [20]. We have analyzed these services and identified their main features and key factors:

In Clarifai, the features available are related to object and colors recognition and detection of explicit content. Price can go up to 2.5 dollars per 1000 images/month. The advantages are the reduced cost and good quality. The disadvantage is the lack of other features.

In Microsoft Cognitive Services (Computer Vision API and Face API), the features available are related to object, colors and text recognition and caption of an image. As these services are in an early stage (released in March 2016), their use is free but with a limit of 2500 images per month. The advantages are the reduced cost, good quality and many appellative features, with a highlight for face identification.

In Google Cloud Vision, the features available are related to object, text, colors, logos and structures recognition and detection of explicit content. The price can go up to 5 dollars per 1000 images/month. The advantages are the reduced cost, good quality and many appellative features.

In Abby Cloud Services, the features available are related to text recognition. Price can go up to 70 dollars per 1000 images/month. The advantages are a good quality and the ability to work with poor quality images. The disadvantages are the lack of other features and the high cost.

From this analysis, we selected Google Cloud Vision and Microsoft Cognitive Services for a deeper analysis, considering the low cost and the additional and adequate features provided.

1. The testbed system

To test the previously selected cloud services, it was created a testbed system as described in the following sections, including the architecture and the system implementation.

1.1. The system architecture

The test system architecture is described in two views, based on the proposal of a multi-view description [21-22].

Fig. 1 illustrates the architecture of the testbed system. It represents a connection view, showing the several elements and their connections; and a process view, illustrating the complete process of assisting a blind user.

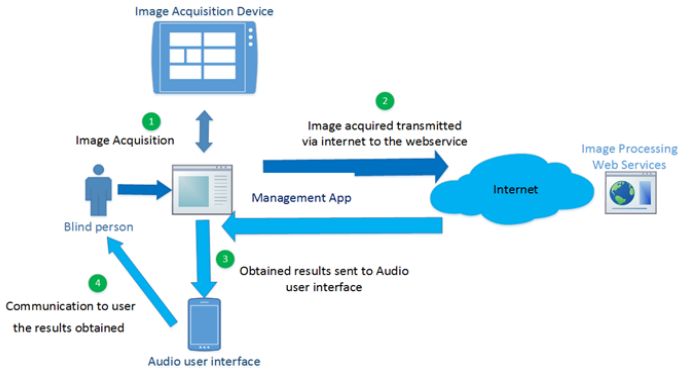
The system comprises the following elements: management app, which coordinates the other elements to execute the user assistance process; image acquisition device, responsible for acquiring the images; image processing service, responsible for processing the images; audio interface, responsible for reporting the analysis results to the blind user.

The user assistance process begins with an image acquisition (1), which is retrieved by the management app and sent for processing in the cloud service (2); then the management app receives the results and sends them to the audio user interface (3); in the final step, the results are transmitted as audio communication to the blind person (4).

Fig. 1. Architecture of the Artificial Vision Module

1.2. The system implementation

The architecture elements were implemented as follows:



- For the acquisition device, it was chosen the Vuzix M100 Smart Glasses, which allows the blind person to easily aim at the objects. This model is the best in the low price range while providing good working features, such as built-in camera and internet connection.
- The cloud services previously selected were: Google Cloud Vision (GCV), featuring object, logo, landmark, text and structure recognition; and Microsoft Cognitive Services (MCS), featuring object and text recognition, the caption of an image and face identification.

For the audio interface, it was chosen a mobile device, the Samsung S7390, which has built inBluetooth and audio output. This device runs an audio user interface application developed specifically to blind people and created by the CE4BLIND project.

- To manage the system it was developed an android app, deployed on the smart glasses. This app could be deployed and run on any other android device with network connectivity. Due to accuracy reasons, it was developed two similar versions of the app, in which one application sends the images to GCV and the other sends the images to MCS. The accuracy reasons are that the increment on latency in having all the features in only one application would compromised the number of attempts needed to process an accurate result. A picture is taken cyclically after 30 seconds, which means that after that time it's considered as a failed attempt.

The system usage starts with the user launches the app on the smart glass, after which

several photos are cyclically taken, processed on the cloud and the results presented to the user by the audio interface, running on the smart phone mobile device.

III. METHODOLOGY

The tests were conducted using the testbed system and taking pictures of a large set of objects, spatially arranged in a quotidian fashion. To have results as good as possible, a mechanism was adopted to only show the realistic results by filtering and selecting only those with a good enough confidence degree.

For each object, it was defined a limit to the number of attempts necessary to reach a correct description of the object. An attempt is defined by the action of taking a picture and checking its accuracy. After the picture was taken, if it was not accompanied by the results of that feature, it is considered a failed attempt. After a failed attempt, the light conditions were modified, the angle of the picture shot was changed or the object position was also changed. If the number of attempts to a specific object was superior to three, the recognition of that object was considered not accurate.

The variables registered were:

- Confidence degree, (only when provided together with the results) representing an estimative of the service efficiency;
- Relative accuracy, calculated as the ratio of the number of accurate results vs. the number of the total shown results.

IV. TESTS AND RESULTS

The tests were realized with the aim to evaluate which features had the best accuracy and were relevant to help the blind people. The features in common in the two external services were text and object recognizing. These two features had common conditions to compare the two cloud services.

1.3. Google Cloud Vision (GCV)

It was tested the features of object, logo, landmark, text and structure recognition.

1.3.1. Object Recognition

It was taken photos of 15 objects and the results were accompanied with a confidence degree. The average of the confidence degree in all results obtained was approximately 68% and the relative accuracy had an average of approximately 66%. Table 2 shows 3 examples of the results.

Table 1. Three results obtained in object recognition using GCV

Object	Results	Confidence (%)		Accuracy (%)		Attempt (max 3)
		Degree	Average	Degree	Relative	
Bible	text	95.0		Accurate	2/2	2
	document	61.1	78.1	Accurate	100%	
Bottle of Water	alcoholic beverage	83.3		Not Accurate	1/3	1
	drink	72.3	72.1	Accurate	33%	
	mobile device	60.7		Not Accurate		
Smart-phone	gadget	86.0		Accurate	2/4	1
	laptop	84.0	75.9	Not Accurate	50%	
	mobile device	69.7		Accurate		
	computer hardware	64.0		Not Accurate		

Overall the results showed that this feature has a reasonable accuracy, and can be used to recognize objects in a supermarket, at home or at work.

1.3.2. Text recognition

It was taken one photo with the app of the first 5 vesicles a bible, written in Portuguese. This feature doesn't provide information about the confidence degree. The text shown in

is the original text, with the length of 288 characters without space.

A Criação
1 No princípio, Deus criou os céus e a terra. 2 A terra era informe e vazia; as trevas cobriram
o abismo e o Espírito de Deus pairava sobre as águas. 3 Deus disse "Faça-se a luz!" E a luz
foi feita. 4 Deus viu que a luz era boa, e separou a luz das trevas. 5 Deus chamou à luz DIA,
e às trevas NOITE. Sobreveio a tarde e depois a manhã: foi o primeiro dia.

Fig. 2. Original text in recognizing text used in GCV and MCS

In the obtained text, 39 characters were missing or wrong. The relative accuracy was calculated by dividing the total characters minus the characters wrong or missing (288-39) by the number of total characters (288). The result is approximately 0.865 (86.5%).

1.3.3. Logo recognition

This feature allows the logo recognition from international brands or well known national brands (tested in well known Portuguese brands). In total it was tested 9 brands, 4 of which were Portuguese. The results were accompanied by the confidence degree.

The relative accuracy was 8 in 9 results (approximately 89%) and the average confidence degree was approximately 21%. The accuracy is very good but the confidence degree doesn´t correlate.

The logo recognition feature was accurate and can be used by blind people at the supermarket identifying popular brands. Fig. shows a demonstration of this feature.

Fig. 3. Demonstration of logo recognition from Google Cloud Vision

1.3.4. Landmark recognition

This feature allows the Landmark recognition from international or national landmarks (the national landmarks tested were from Portugal). It was chosen 4 international landmarks and 3 national landmarks. The images were taken at showing the landmarks onan LCD screen. The **Error! Reference source not found.** shows the results gathered in this feature and were accompanied with the confidence degree.

Landmarks	Results	Confidence degree	Attempts Max 3
Eiffel Tower	---	---	4
Big Ben and Houses of Parliament	Houses of Parliament	45.9	3
Statue of Liberty	Statue of Liberty	36.3	3
White House	White House	41.9	1
Mosteiro dos Jerónimos	“Jerónimos Monastery”	24.3	1
Torre de Belém	“Belém Tower”	88.9	1
Santuário de Nossa Senhora de Fátima	“Fatima Sanctuary”	61.9	2

Table 3 All the results obtain in landmark recognition using GCV

From these results, 6 in 7 were accurate (relative accuracy approximately 85.7%). The average of the confidence degree was 42.7%. This feature can be used to blind people recognize famous buildings nationally or internationally.

1.4. Microsoft Cognitive Services (MCS)

It was tested the features of object and text recognition, the caption of an image and face identification. The Microsoft Cognitive Services needed an extra module (available all code in the Github Microsoft Cognitive Services) for training the service to identify familiar images for the feature of face identification.

1.4.1. Object Recognition

Were taken photos to 15 objects and the results were accompanied with a confidence degree, the average of the confidence degree in all obtained results was 74% and the relative accuracy had an average ofapproximately 92.5%.
The results shown that the confidence degree underestimates the accuracy of the recognition, for that, it shouldn´t be applied any filter because it would filter results that were accurate.

1.4.2. Text recognition

Were taken three photos with the app of the first 5 vesicles in the Portuguese Bible. It needed three attempts because the first two results were very weak. This feature doesn't provide the confidence degree information. The text in is the original text, with the length of 288 characters without space.

In the obtained text, 65 characters were missing or wrong. The relative accuracy was calculated from the total characters minus the characters wrong or missing (288-65) by the number of total characters (288), the result is approximately 0.774

Table 2 shows 3 results of caption using MCS.

Table 2. Three results obtained in caption (description of a scenario) using MCS

Object	Results	Relative Accuracy	Confidence (%)	Attempt (max 3)
Bible	"a book sitting on a desk "	2/2 100%	58.4	3
Bottle of Water	"a bottle of wine"	1/2 50%	28.4	1
Smartphone	"a desk with a cellphone"	2/2 100%	76.2	1

Overall, the results were:

Total accurate (relative accuracy of 100%) – 9 results shown (confidence degree between 7.9% and 92.2%)

Satisfying accurate (relative accuracy between >0% and < 100% / it as at least one characteristic accurate) - 5 results shown (confidence degree between 28.4% and 86.9%)

Not accurate (passed the max of 3 attempts/ relative accuracy of 0%) – 1 result shown (confidence degree of 48.9%)

The average of relative accuracy was 84.4%, which shows that the results shown were accurate. The average of the confidence degree was 54.7%. These results were reasonable accurate, but the confidence degree was very inconsistent, thus it can´t be used to filter the results. This feature can be useful to give the blind people a context of the surrounding environment.

1.4.4. Face identification

This feature provides a face identification service requiring some previous service training. The app developed for the testbed system has the option to allow the user to add a person to the service and train the identification of many faces. It is recommended to train the service at least with 3 images of different angles of the target person. The results were accompanied by the confidence degree. The tests were realized with 3 people,

(77.4%). This feature can be used to aid blind people to identify the content of a menu from a restaurant or the main letters of some products. This feature had difficulties to recognize text with small characters.

1.4.3. Caption (description)

This feature does a description of an image, based on object recognition, and it was tested on 15 objects. The results were accompanied with a confidence degree. The criterion chosen was relative accuracy but this time with a different formula, instead of calculated as the ratio of the number of accurate results versus the number of the total shown results is calculated as the ratio of the number of accurate characteristics versus the number of the total shown characteristics. The reason is that for each image taken, most of the time only one result came, and that result had many characteristics. For example, it was taken a picture to a bottle of water, and the results came were: "a bottle of wine". From this sentence, we can identify two characteristics: a bottle and wine. The bottle is accurate and the wine is not accurate.

previously trained the service with 3 faces for each person, the confidence degree average was approximately 84%. The relative accuracy was 100%. The results shown were excellent. It can be very useful to blind people recognize friends and familiars. It is recommended that other person than a blind people take pictures to train this service with more accuracy, so that this feature works better.

V. CONCLUSION AND DISCUSSION

Overall the results were good and we could use the services to accurately identify objects and text. Both services, GCV and MCS are very promising for usage as elements of future systems to help blind people.

The results obtained show that Google API had better accuracy at text recognizing and Microsoft had better accuracy at object recognizing. These two features were common at both APIs. The other features all had good accuracy and important relevance, but the highlight goes to face identification from Microsoft API that enables the blind people to identify known faces.

The tested features, common to GMC and MCS, were:

- Text recognition, on which for both APIs, the relative accuracy was great. The API from Google had the 86.5 % in only one attempt from relative accuracy and the Microsoft had, in three attempts, 77.4% relative accuracy. The text was the same, with these results we can conclude that the Google API is better in recognizing text, useful to read menus in a restaurant, read bills or even recognizing content in many products at a supermarket.
- Object recognition, on which both APIs, the relative accuracy was good, with the spotlight gone to the API of Microsoft with better relative accuracy comparing to the Google API. The API of Microsoft should be used to do object recognizing, and could be used by blind people at the supermarket (identifying type of products), at home (identifying home appliances), at work (identifying electronic devices) and in nature (identifying plants…).

The features, unique to Google Cloud Vision:

- The logo recognition feature was accurate and can be used by blind people, e.g., at the supermarket to identifying popular products.
- The landmark recognition feature was also accurate and can be used to identify national and international landmarks, especially popular buildings, and monuments.

The feature unique to Microsoft Cognitive Services:

- The caption (description) feature had a reasonable accuracy and can be used to give a blind people a reasonable perception of the contextual environment where he is inserted.
- The face identification had good accuracy and can be used to identify familiars. For better accuracy, it should be another person beside the blind person to train the service and have multiple images of each person in the service.

FUTURE WORK

The tests were very promising and the usage of artificial vision based on cloud services is on its infancy, so future work will be of great importance. We expect to research the integration of these services on systems to assist blind people in an autonomous way, by independently managing the discovery, evaluation and usage of the most adequate online cloud services for each usage context.

ACKNOWLEDGEMENTS

This work was supported by the Project “Nano STIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016" financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

REFERENCES

[1] WHO, 2016. “WHO | Blindness: Vision 2020 - The Global Initiative for the Elimination of Avoidable Blindness” http://www.who.int/mediacentre/factsheets/fs214/en/

[2] R Manduchi, Mobile Vision as Assistive Technology for the Blind: An Experimental Study, 13th International Conference on Computers Helping People with Special Needs (ICCHP) (2012).

[3] V Santos, L Amaral, H Mamede, Information Systems Planning - How to enhance creativity?, CENTERIS'2011 - Conference on Enterprise Information Systems

[4] A Reis, D Paulino, H Paredes, J Barroso, Using Intelligent Personal Assistants to Strengthen the Elderlies’ Social Bonds, Universal Access in Human–Computer Interaction, Human and Technological Environments 01 (2017) 593-602.

[5] B Gonçalves, T Rocha, A Reis, J Barroso, AppVox: An Application to Assist People with Speech Impairments in Their Speech Therapy Sessions, Recent Advances in Information Systems and Technologies 03 (2017) 581-591.

[6] W Xindong, Z Xingquan, W Gong-Qing, D Wei, Data Mining with Big Data, IEEE Transactions on Knowledge and Data Engineering 26 (2014).

[7] P Domingos, A Few Useful Things to Know about Machine Learning, Communications of the ACM 55 (2012) 78-87.

[8] Y Bengio, Deep Learning of Representations: Looking Forward, Lecture Notes in Computer Science 7978 (2013) 1-37.

[9] K Hallman, Artificial Intelligence, Zygotes, and Free Will, Res Cogitans 6 (2015) 2155-4838.

[10] R Kamberov, C Granell, V Santos, Sociology Paradigms for Dynamic Integration of Devices into a Context-Aware System, Journal of Information Systems Engineering & Management 2 (2017) 2468-4376.

[11] INESC TEC, 2016. “ Olha o CE4BLIND”, https://www.inesctec.pt/csig/noticias-eventos/nos-na-imprensa/olha-o-ce4blind-quer-dizer-tecnologias-para-aumentar-a-autonomia-dos-invisuais plural-singular/

[12] T Rocha, H Fernandes, A Reis, D Paulino, H Paredes, J Barroso, Assistive Platforms for the Visual Impaired: Bridging the Gap with the General Public, Recent Advances in Information Systems and Technologies 03 (2017) 602-608.

[13] INESC TEC, 2016. “INESC TEC”, https://www.inesctec.pt/

[14] Reis A., Barroso I., Monteiro M., Khanal S., Rodrigues V., Filipe V., Paredes H., Barroso J., 2017. “Designing Autonomous Systems Interactions with Elderly People”. Universal Access in Human–Computer Interaction. Human and Technological Environments, 01/2017: pages 603-611; , ISBN: 978-3-319-58699-1, DOI:10.1007/978-3-319-58700-4_49

[15] T Adão, M Luís, H Paredes, J Barroso, Navigation module of Blavigator prototype, World Automation Congress 2012, http://ieeexplore.ieee.org/document/6320942/

[16] Adão T., 2011. “Módulo de Navegação para Cegos (Smartvision)”, UTAD, http://hdl.handle.net/10348/2094

[17] Clarifai, 2016. “Clarifai”, https://www.clarifai.com/technology

[18] Microsoft, 2016. “Microsoft Cognitive Services - APIs”, https://www.microsoft.com/cognitive-services/en-us/apis

[19] Google, 2016. “Vision API - Image Content Analysis | Google Cloud Platforms” https://cloud.google.com/vision/

[20] ABBYY, 2016. “Abby Cloud OCR SDK” ,http://ocrsdk.com/

[21] P. B. Kruchten, The 4+1 View Model of architecture, IEEE Software 12 (1995) 42–50.

[22] D. E. Perry, A. L. Wolf, The three views (processing, data, and connection) in software architecture by Perry and Wolf, ACM SIGSOFT Software Engineering Notes 17 (1992) 40–52