Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 91–96
DOI: 10.15439/2019F345
ISSN 2300-5963 ACSIS, Vol. 19

Integrating Computer Vision and Natural Language Processing to Guide Blind Movements

Lenard Nkalubo
Department of Computer Science
Kyambogo University

***Abstract*—Vision is the most essential sense for human beings. Vision impairment is one of the most problems faced by the elderly. Blindness is a state of lacking the visual perception due to physiological or neurological factors. This paper presents a detailed systematic and critical review that explores the available literature and outlines the research efforts that have been made in relation to movements of the blind and proposes an integrated guidance system involving computer vision and natural language processing. An advanced Smartphone equipped with vision, language and intelligence capabilities is attached to the blind person in order to capture surrounding images and is then connected to a central server programmed with a faster region convolutional neural network algorithm and an image detection algorithm to recognize images and multiple obstacles. The server sends the results back to the Smartphone which are then converted into speech for the blind person's guidance.**

***Index Terms*—Computer vision, Smartphone-based, Faster CNN algorithm, visually impaired, natural language processing.**

I. Introduction

TRADITIONALLY, movements of the blind are guided by a walking stick. As technologies improve, smart walking sticks have been explored by embedding sensors on the walking sticks. Other attempts have also been tried with the use of electronic travel aids [1], electronic orientation aids (EOAs) [13] and position locator devices (PLDs) [10]. Despite all the efforts undertaken to solve the movement of the blind, it remains challenging and requires more research endeavors [2].

This paper gives the state of the art and outlines the research efforts in relation to the techniques involved in the movement of the blind.

The rest of this paper is organized as follows: section 2 discusses the research motivation or concern, section 3 provides the current literature about the techniques involved in solving the problem, section 4 gives the methodology, section 5 gives the intended research product and section 6 gives the conclusion.

II. Research Motivation

Globally, it is estimated that approximately 1.3 billion people live with some form of distance or near vision impairment. With regards to distance vision, 188.5 million have mild vision impairment, 217 million have moderate to severe vision impairment, and 36 million people are blind [3]. With regards to near vision, 826 million people live with a near vision impairment. The study carried out in [3] identified 288 studies of 3,983,541 participants contributing data from 98 countries. Among the global population with moderate or severe vision impairment in 2015 (216·6 million [80% uncertainty interval 98·5 million to 359·1 million]), the leading causes were uncorrected refractive error (116·3 million [49·4 million to 202·1 million]), cataract (52·6 million [18·2 million to 109·6 million]), age-related macular degeneration (8·4 million [0·9 million to 29·5 million]), glaucoma (4·0 million [0·6 million to 13·3 million]), and diabetic retinopathy (2·6 million [0·2 million to 9·9 million]) [3].

Furthermore, 81 percent of people with vision impairment are aged 50 and above years. Apart from age, other causes of vision impairment have been found to be cataracts, glaucoma, diabetic retinopathy, and uncorrected refractive errors [6]. The number of people affected by the common causes of vision loss has increased substantially as the population increases and ages. Preventable vision loss due to cataract (reversible with surgery) and refractive error (reversible with spectacle correction) continue to cause most cases of blindness and moderate or severe vision impairment in adults aged 50 years and older. A large scale-up of eye care provision to cope with the increasing numbers is needed to address avoidable vision loss. [3]

Arising from this statistics, it is clear that the problem of vision impairment cannot be addressed fully from the medical perspective. We have to explore other alternatives that support those already in existance since blindness or vision impairment is function of age which puts the aging persons at high risk of becoming visually impaired.

A. Research Objective

The general objective of the paper is to identify current literature, current research efforts in solving the problem of the movements for the blind or unpaired people while proposing the latest solution which emerges from current technological advancements in artificial intelligence with the integration of computer vision and NLP. In a more specific way this would be done by finding out the current strength and weakness of the blind movement solutions, identifying computer vision Algorithms and NLP capabilities especially deep learning CNN algorithms and latest smart phones supporting natural language capabilities.

B. Research Question

The general research question would be to find out how can the current research efforts of solving the vision impairment problem be analyzed? How can new trends in technology like computer vision and Natural language processing be used in solving the problem?

In more specific terms the research questions would address: What are the strengths and weaknesses of the current blind movement solution systems? What are the requirements of the integrated computer vision and Natural language processing proposed solution? How can a faster CNN-algorithm be implemented with language capabilities to provide a solution for blind movements?

In summary, according to the implementations in previous studies, assistive devices for navigation for visually impaired people still focus on location and distance sensing and alerting users on the types of obstacles in front of them and their surroundings. Therefore, the practicability of such assistive devices is very low due to the cost and vulnerability to damage from the sun and rain [9]. Therefore, this paper addresses those obstacles and proposes a mobile navigation system for visually impaired people; this system employs an advanced Smartphone and with deep learning algorithms to recognize various obstacles and is not limited to indoor or outdoor environments.

The Google 3xL Smartphone released in November 2018 whose Image processor and sensor is shown in (Fig 1). is equipped with machine learning capabilities like text-speech, image recognition, voice processing and facial recognition and has Google global positioning system features (A-GPS, GLONASS, BDS, and GALILEO). The Smartphone has Google applications like the Google's cloud, maps, lens, and assistant. This latest advanced Smartphone addresses gaps existing in the old non mobile systems and makes the study address the research gap in previous implementations. The study, investigation, analysis, design and implementation of these new technologies will squarely bridge the research gap and contribute a new knowledge base towards bridging computer vision and natural language processing.

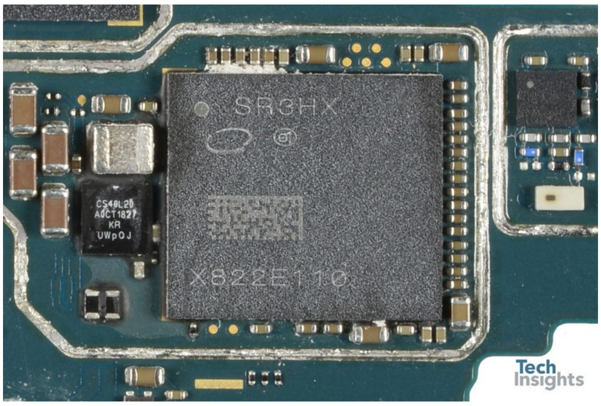

Fig. 1. Google – Pixel - 3XL –Image Processor and Sensor

III. Literature Review

A. Existing Technologies

A smart blind guidance device has been proposed by [1] which uses infrared sensors and includes a small hand stick along with a wearable head set. Another system known as Sonic Path Finder shown in (Fig 2). by [4] which works based on ultrasound transmission and reception but it is not a hand held system, it is attached on the head of the user but is unable to provide the accurate path and the position of an obstacle. It is a secondary mobility aid for use by people with vision impairment. It is not suitable for anyone who does not have primary mobility skills. It is designed for use out-of-door in conjunction with a cane, guide dog or residual vision [4].

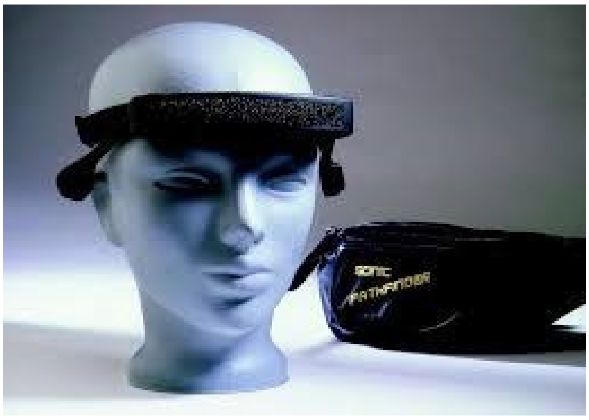

Fig. 2. Sonic Path Finder

ETAs are general assistant devices to help visually impaired people avoid obstacles [5]. Microsoft Kinect shown in (Fig. 3) is usually used as the main recognition hardware in such systems [11]. However, Microsoft Kinect cannot be used in environments with strong light. Moreover, it can determine only the presence of obstacles ahead [8] or recognize a few types of obstacles in few related studies [7]. In general camera recognition systems are designed to recognize tactile or obstacle images.

Fig. 3. Microsoft Kinetic Sensor

A combination of a camera and other multiple sensors is usually used to get more information to draw the shapes of passageway and obstacles [7]. Thus these systems may provide a guiding service and a recognition result out of a few types of obstacles. The drawback of EOAs is that they need more complex computing to hardly be realized as a real-time and lightweight guiding device. PLDs are used to determine the precise position of its holder such as devices that use global positioning system (GPS) and geographic information system (GIS) technologies [2].

B. *Computer Vision*

Computer Vision (CV) tasks can be summarized by the concept of 3Rs [12], which are reconstruction, recognition, and reorganization. Reconstruction involves estimating the three-dimensional (3D) scene that gave rise to a particular visual image. This representation is shown in (Fig 4). It can be accomplished using a variety of processes incorporating information from multiple views, shading, texture, or direct depth sensors. Reconstruction process results in a 3D model, such as point clouds or depth images. Some examples for reconstruction tasks are Structure from Motion, scene reconstruction, and shape from shading. Recognition involves both 2D problems (like handwritten recognition, face recognition, scene recognition, or object recognition), and 3D problems (like 3D object recognition from point clouds which assists in robotics manipulation). Recognition results in assigning labels to objects in the image. Reorganization involves bottom-up vision: segmentation of the raw pixels into groups that represent the structure of the image. Reorganization tasks range from low-level vision like edge, contour, and corner detection, intrinsic images, and texture segmentation to high-level tasks like semantic segmentation [15], which has an overlapping contribution to recognition tasks. A scene can be segmented based on low-level vision] or high-level information like shadow segmentation that utilizes class information.

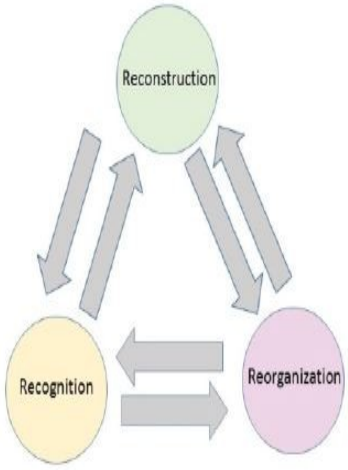

Fig.4 The 3Rs in computer vision (Malik et al. 2016)[12], which are reconstruction, reorganization and recognition.

C. *Natural Language Processing (NLP)*

Following the Vauquois triangle for machine translation shown in (Fig 5).[17], Natural Language Processing (NLP) tasks can be summarized into concepts ranging from syntax to semantics and to pragmatics at the top level to achieve communication. Syntax includes morphology (the study of word forms) and compositionality (the composition of smaller language units like words to larger units like phrases or sentences). Semantics is the study of meaning, including finding relations between words, phrases, sentences or discourse. Pragmatics studies how meaning changes in the presence of a specific context. For instance, an ironic sentence cannot be correctly interpreted without any side information that indicates the indirectness in the speaker's intention.

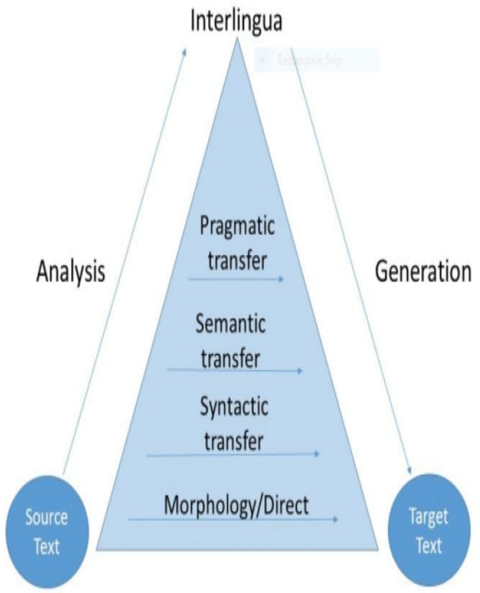

Fig 5. The Vauquois triangle for machine translation (Vauquois, 1968) [17].

Ambiguity in language interpretation a main obstacle for an intelligent system to overcome and achieve language understanding. Some complex tasks in natural language processing include machine translation, information extraction, dialog interface, question answering, parsing, and summarization. There is always meaning lost when translating between one language and another. When "translating" between the low-level pixels or contours of an image and a high level description in word labels or sentences, there is a wide chasm to be crossed. Bridging the Semantic gap means building a bridge from visual data to language data like words or phrases.

D. *Conceptual Framework*

As a general framework shown in (Fig. 6), most methods in image captioning are trying to either model language information as another layer on top or jointly model language and vision simultaneously by a carefully designed loss function or algorithm. These systems consider structural multimodal input and create structural output in contrast to the traditional system.

[15] Unifies language and vision for robotics again by bridging visual, language, speech, and control data for a forklift robot. Their robot can recognize objects based on one example using one-shot visual memory. Its natural language interface works by speech processing or pen gestures.

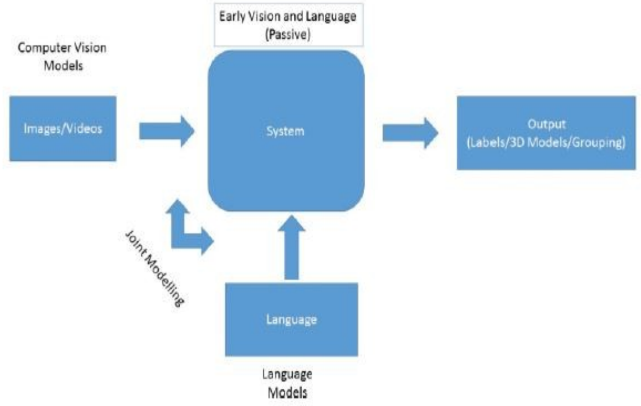

Fig.6. The early vision and language conceptual framework which passively utilizes high-level language information as an additional context. memory.

[18] provides a survey for verbal and nonverbal human-robot interaction.

A video can be described by a sentence or a discourse that is a structured set of sentences that tells a specific story. A sentence prior can be learned from web-scale corpora to bias the model and penalize unlikely combinations of actors/actions/objects [19]. It links the compositional structure of languages and the compositional structure of video events using natural language semantics and three essential computer vision tasks which are tracking, object detection, and event recognition. These three tasks are done simultaneously using a single cost function including the attention mechanism to focus on the most salient event to produce the best sentence description for activity recognition.

From a natural language processing perspective, the sentence tracker utilizes lexical semantics and contains the information of “who did what to whom, and where and how they did it” [20]. An object is described as a noun phrase; the observed action is described as a verb; object properties are described as adjectives; spatial relations between objects are described by prepositions; and the event characteristics are represented as prepositions and adverbs. The system has a predefined vocabulary and a sentence is composed using a set of predefined grammars.

The sentence tracker can be divided into two subsystems [20]. The first subsystem consists of three steps. First, object detection is performed with a high-recall setting. Second, tracking with forward projection to increase precision is performed. Third, the optimal set of detections is chosen using dynamic programming with the Viterbi algorithm, outputting a result consistent with the optical flow. For the second subsystem, events are recognized using HMMs, also computed using the Viterbi algorithm. The unified objective function from the final step of the first subsystem and the second subsystem can be merged, since both are based on HMMs.

Recent approaches deploy powerful deep-learning frameworks to model both image and word sequences. These approaches can support a larger vocabulary than other methods that have a small set of predefined vocabulary [21]. Similar to image captioning, [21] combines a sequence of CNN and another two sequences of LSTM to generate sentence description from video. AlexNet is deployed as a pretrained CNN model, and the output features are mean-pooled before feeding to the LSTM sentence decoding layer. This work is inspired by [22], which uses CRF to extract image features for the intermediate representation for an LSTM. [21] makes an improvement to [22] that discards the temporal information and models the image frames as a bag-of-images.

E. Fast CNN (Deep Learning)

There are many attempts in many benchmarks in open competitions to design a better architecture of CNNs. Some notable architectures are AlexNet, GoogLeNet, VGGnet and ResNet [16]. The main insight from these models is that deeper models are better for classification. Based on these models for recognition, more models are proposed for other computer vision tasks. For example, R-CNN or Fast R-CNN have been proposed for object detection. Another widely used architecture is FCN for semantic segmentation. It is a fully convolutional neural network that can perform pixel wise labeling. The idea delves further into a deeper problem of structured prediction when recurrent neural networks can be seen as a generic sequence model like CRFs.

IV. Research Methodology

A. Research Philosophy

The philosophical stance of the research is highlighted in the research onion’s outermost layer. According to (Saunders et al., 2012) [16] there are four different philosophical branches that define the presence of a research entity; the first is positivism, the second is realism, the third is interpretivism and the fourth is pragmatism.

On the basis of empirical evidences and prior theories on brand management, brand choice frameworks will be examined in the current research. Positivism is mainly based on strong observation and forecasting outcomes similar to a laboratory scientist, with the aim of obtaining law-like generalizations for ascertaining cause and effect.

The researchers who adopt this approach underline the use of ‘scientific method’ for proposing and testing theories that have highly measurable and structured data, wherein the values of the researcher do not influence the research. Thus, this approach supports large samples of quantitative data which is analysed along with statistical testing of hypothesis. Such an approach helps test a theory, confirm a theory or revise a theory based on the analysis of the existing data.

This paper proposes positivism.

B. Research Design

The research design is experimental. The proposed navigation system employs a Google 3xL Smartphone. It will be used to continually capture images of the environment in front of a user and perform image processing and object identification to inform the user of the image results. The Uganda National Institute for Special Education (UNISE), Kyambogo University will be used as an experimental scene. UNISE is a national institute for students with special needs especially disabled students. On the other hand, the specification of Linux server hardware is a modern personal computer equipped with an i7 central processing unit (CPU) 64-bit i7 Intel/AMD-based PC and 4 gigabytes (GB) of RAM or higher and a graphics processing unit NVIDIA GeForce GTX 1050 GPU (or higher)to execute deep computing modules which are based on the faster region convolutional neural network (Faster R-CNN) algorithm.

C. *Software Design*

The software design of our proposed system involves the feature recognition, deep recognition, and direction and distance modules. There will be also a mobile application for interfacing the above modules. These modules and the application will be developed using Open CV, Java and Python computer programming languages.

D. *Experimental Data collections*

The experimental tests will be carried out on various candidates in the University. The tests will based on gender, age, degree of visual impairment, past experience on the use of similar or related equipment, literacy level, type of obstacle during movement and duration of time during the use of the experiment. Practically it is proposed that Women, children and the elderly may be given priority. Interviews may be carried out in assessing the performance, accuracy and speed of the equipment. The findings will be used to improve the performance and use of the equipment. About the experiment process, the participants will be required to turn the camera lens of smartphone to the front side and walk through institute campus from a main building through the parking towards the institute main gate. Each experiment will be arranged in the morning, noon and evening time. Before the experiment is carried out, a training session involving stakeholders can be arranged on how to use the equipment and precautions noted. This process would allow each participant to know the usage of the system. The selected participants may include five female and three male visually impaired students with a range of 17-25 years and also four old people like two females aged between 50-60 years and two males aged between 60-70 years. To obtain the degree of accuracy the degree of visual impairment should be similar. It should also be noted that this system will work for people with a hearing sense. The blind and deaf will not be managed by the proposed system. It should also be noted that the system will work for virgin disabled people who have never used any other system.

V. Research Product

The intended outcome of this research paper is a solution which solves the problem of movement for the blind people. This solution is a Smartphone-based guiding system for solving the navigation problems for visually impaired people and achieving obstacle avoidance to enable visually impaired people to travel smoothly from a beginning point to a destination with greater awareness of their surroundings. Blind people find it hard to walk through busy roads and travel new places and so the guided Smartphone will become their daily companion. This product is simple, cheap, user friendly and it is designed and implemented to improve the mobility of both blind and visually impaired people in a specific area.

VI. Conclusion

This paper proposed the development of a user-friendly guidance system for the visually impaired people. This system involves an advanced Smartphone and a Linux server connected and processed using a deep learning algorithm for image recognition. When the system is in use, the smart phone would continuously transmits images of the scene in front of the user to a server through using a 4G technology or a Wi-Fi network. Subsequently, the server performs the recognition process and the final results are transmitted back to the smart phone. The system would provide the user with obstacle track and avoidance information through voice notifications.

In the future, to provide information on more types of obstacles and more accurate recognition, a broader range of obstacle images and a high-end server equipped with a more powerful graphics processing unit could be used to increase the number of recognition categories and the recognition rate. The system is recommended for the blind people with a hearing sense and suitable for less traffic environments like universities, prisons and hospitals not busy city roads.

Acknowledgments

This research was partly supported by Kyambogo University Staff development committee under the Staff development vote. The lectures, guidance and comments from Prof Josephine Nabukenya of the department of Information systems, Makerere University, college of computing and informatics that helped the writer to furnish up the paper.

I wish also to acknowledge my fellow PhD students at Makerere students in the departments of computer science-Artificial intelligence lab, and department of computer networks for the various seminars conducted weekly on Thursdays which motivated me to carry out these research ideas.

Finally I wish to thank my reviewers who were appointed by the Federated Conference on Computer Science and Information Systems (FedCSIS) (DS-RAIT'19) Program Committee who reviewed this paper when it was submitted. I found your comments, directions and criticisms so important to furnish up this final revised position paper.

References

[1] A. S .Al-Fahoum., H. B .Al-Hmoud., and A. A.Al-Fraihat, "*A smart in-frared microcontroller-based blind guidance system*", Active and Passive, Electronic Components, vol. 2013

[2] V. Adagale and S. Mahajan,*" Route Guidance System for Blind People Using GPS"* and GSM. IJEETC ,4,16–21, 2015

[3] R. R. A. Bourne , S. R. Flaxman, and T. Braithwaite "*Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis".* Lancet Glob Health. 5: e888-e897, 2018

[4] S. Chaurasi and V.N. Kavitha, "*An Electronic Walking Stick for Blinds, in Information Communication and Embedded Systems*" (ICICES), 2014 International Conference on. IEEE, 2014, pp. 1–5

[5] V. Filipe, F. Fernandes, H. Fernandes, A. Sousa, H. Paredes, J. Barroso, *"Blind navigation support ystem based on Microsoft Kinect. In Conf. Proceedings of the 2012 International Conference on Software Development for Enhancing Accessibility and Fighting Info-Exclusion"* (DSAI), Douro, Portugal, pp. 94–101. 2012

[6] T. R. Fricke, N. Tahhan, E. Resnikoff,, A. Burnett, S. M. Ho, T. Naduvilath, K. S. Naidoo. "*Global Prevalence of Presbyopia and Vision Impairment from Uncorrected Presbyopia: Systematic Review, Meta-analysis, and Modelling. Ophthalmology"*. 2018 Oct; 125(10):1492-1499. doi: 10.1016/j.ophtha.2018.04.013

[7] V. N Hoang,T H. Nguyen, T. L. Le, T. H Tran,T. P Vuong, N. Vuillerme, "*Obstacle detection and warning system for visually impaired people based on electrode matrix and mobile Kinect"*. Vietnam J. Comput. Sci.,4, 71–83. 2017

[8] H. C. Huang, C. T. Hsieh, C. H Yeh, "*An Indoor Obstacle Detection System Using Depth Information and Region Growth. Sensors"* 2015, 15, 27116–27141.

[9] S. L Joseph., J. Xiao., X. Zhang., B. Chawda, K. Narang., N. Rajput., S. Mehta., L.V. Subramaniam, "*Being Aware of the World: Toward Using Social Media to Support the Blind with Navigation"*. IEEE Trans. Hum. Mach. Syst. 45, 399–405. 2015

[10] A. M Kassima., T. Yasunoa., M. S. M. Arasb., A. Z Shukorb, H. I. Jaafarb., M. F. Baharomb., F. A. Jafarb,. "*Vision Based of Tactile Paving Detection in Navigation System for Blind Person"*. J. Teknol. (Sci. Eng.) , 77, 25–32. 2015

[11] S. Mann, J. Huang, R. Janzen, R. Lo, R. Ramoersadm, V. Chen, A. Doha, *"Blind Navigation with a Wearable Range Camera and Vibrotactile Helmet"*. In Conf. Proceedings of the 19th ACM International Conference on Multimedia, Scottsdale, AZ, USA, pp. 1325–1328. 2011

[12] J. Malik, A. Arbeláez, C. Joao, F. Katerina., G. Ross, G. Georgia, S. Gupta, H. Bharath, A. Kar., and T. Shubhami. "*The three Rs of computer vision: Recognition, reconstruction and reorganization" Pattern Recogn*. Lett. 72 , 4–14. 2016

[13] A. Pereira., N. Nunesa., D. Vieiraa., N. Costaa., H. Fernandesc., J. Barroso, "*Blind Guide: An ultrasound sensor-based body area network for guiding blind people*." Procedia Comput. Sci., 67, 403–408. 2015

[14] M. Saunders., P. Lewis and A. Thornhill, "*Research Methods for Business Students"*. Pearson Education Ltd., Harlow. 2012

[15] R. Socher , C.D. Manning, and A.Y. Ng, " *Parsing natural scenes and natural language with recursive neural networks*." In Conf. 28th International Conference on Machine Learning (ICML-11). 129–136. 2011

[16] C. Szegedy, L. Wei , J. Yangqing, S. Pierre, R. Scott., A. Dragomir., E. Dumitru, V. Vincent., R. Andrew "*Going deeper with Convolutions*." in IEEE Conference on Computer Vision and Pattern Recognition, 1–9, 2015.

[17] B. Vauquois, "*A Survey of Formal Grammars and Algorithms for Recognition and Transformation in Mechanical Translation*." In Conf. Proceedings of IFIP Congress, 1114–1122. Edinburgh. 1968

[18] M. R Walter, M. Antone, E. Chuangsuwanich, A. Correa, R. Davi, L. Fletcher, E. Frazzoli, Y. Friedman, H. P. Jonathan, H. Jeong, S. Karaman, B. Luders, J. R. Glass, "*A situationally aware voice-commandable robotic forklift working alongside people in unstructured outdoor environments*." J.Field Robot. 32, 4, 590–628. 2015. DOI: 10.1002/rob.21539

[19] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, Kate Saenko, and T. Darrell. 2015. *Long-term recurrent convolutional networks for visual recognition and description.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2625–2634.

[20] H. Yu, N. Siddharth, A. Barbu, and J. M. Siskind. *A compositional framework for grounding language inference, generation, and acquisition in video*. J. Artif. Intell. Res. (2015), 601–713.

[21] H. Xu, S. Venugopalan, V. Ramanishka, M. Rohrbach, and K. Saenko. 20. *A multi-scale multiple instance video description network*. arXiv preprint arXiv:1505.05914 (2015).

[22] S. Venugopalan, H. Xu, J. D.Marcus Rohrbach, R. Mooney, and K. Saenko. *Translating videos to natural language using deep recurrent neural networks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015