An Enhanced Intelligent Agent with Image Description Generation

Ben Fielding, Philip Kinghorn, Kamlesh Mistry, and Li Zhang

Department of Computer Science and Digital Technologies, Facutly of Engineering and Environment, Northumbria University, Newcastle, NE1 8ST, United Kingdom
{ben.fielding, philip.kinghorn, kamlesh.mistry, li.zhang (corr. author)}@northumbria.ac.uk

Abstract. In this paper, we present an Embodied Conversational Agent (ECA) enriched with automatic image understanding, using vision data derived from state-of-the-art machine learning techniques for the advancement of autonomous interaction with the elderly or infirm. The agent is developed to conduct health and emotion well-being monitoring for the elderly. It is not only able to conduct question-answering via speech-based interaction, but also able to provide analysis of the user's surroundings, company, emotional states, hazards and fall actions via visual data using deep learning techniques. The agent is accessible from a web browser and can be communicated with via voice means, with a webcam required for the visual analysis functionality. The system has been evaluated with diverse real-life images to prove its efficiency.

Keywords: Intelligent conversational agent · image description generation · human agent interaction.

1 Introduction

We propose a system to assist with the day-to-day care of the elderly by attempting to emulate some of the human contact they receive throughout the day. By providing an elderly person with a non-human, Embodied Conversational Agent (ECA) companion, we improve access to information, and provide accident prevention and reaction functionality. Evidence shows that the social interaction provided by ECAs has a positive effect on the interacting user, whilst information provided through conversation with an ECA is more easily absorbed and understood [1]. The interface of the proposed system is illustrated in Fig. 1.

To achieve these goals, we incorporate a number of unique computer vision techniques together, along with an approachable 3D humanoid avatar interface with real-time chat functionality. Intelligent Chat, proposed in this research, provides real-time visual, audial and oral conversation with a number of additional features tailored to the needs of elderly users. It provides answers to queries on any subject using integration with the popular online encyclopedia Wikipedia. Intelligent Chat provides live feedback of the emotional state of the user using current vision-based facial emotion recognition techniques applied to the user's webcam, all performed in the browser. Moreover, the user's environment is monitored by the system using state-of-the-art deep learning computer vision techniques embedded in a central server system, providing analysis of the overall scene, objects, potential hazards around the user, and alerting in the event of a fall. This vision-based analysis, performed on the central server, is used to enhance the conversation with the user, thereby providing an engaging and life-like companionship experience.

The main contribution of this research is the addition of deep learning based image description generation functionality to an ECA, allowing the agent to provide health surveillance. It is one of few pioneer systems in incorporating vision-based analysis to enable more autonomous agent behaviours to enhance user experience. Moreover, existing research on image description generation tends to employ a holistic method, thereby encountering limitations in handling cross-domain images. Compared with the existing related research, our work employs a local region-based approach, thus has great robustness in dealing with cross-domain images that the system has not been trained upon (e.g. images with fall and hazard situations).

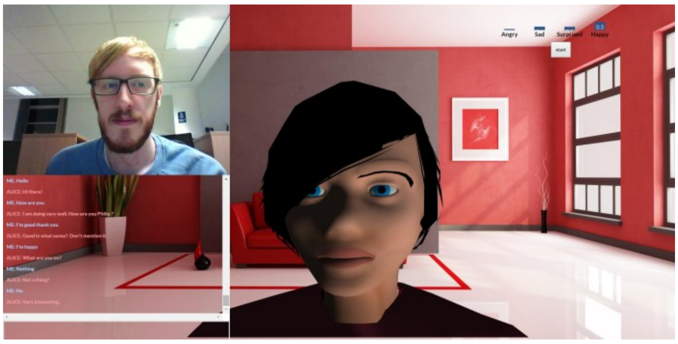

ME: Hello
ALICE: Hi there!
ME: How are you
ALICE: I am doing very well. How are you Philip?
ME: I'm good thank you
ALICE: Good in what sense?. Don't mention it.
ME: I'm happy
ALICE: What are you on?
ME: Nothing
ALICE: Not a thing?
ME: No
ALICE: Very interesting.

Fig. 1. The interface of the proposed Intelligent Chat system

2 Related Work

Care of elderly and infirm patients often requires round-the-clock observation to prevent or react to accidents that can result in physical injury. This constant presence of a healthcare provider is often impossible to achieve due to the number of patients greatly outweighing the number of carers. Therefore, vision-based health monitoring systems using computer vision techniques are required. Image description generation techniques have gained intensive research attention recently to benefit such applications. Some of such developments are discussed below.

2.1 Image Description Generation

Image description generation has been the focus of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) since its conception in 2010. The ILSVRC provides training and testing datasets of images and appropriate labels and presents a number of challenges such as object detection, localization, and scene classification. Prior to 2012, the ILSVRC witnessed a variety of classification techniques used with varying effectiveness. Krizhevsky et al. [2] proved, in 2012, that Deep Convolutional Neural Networks (CNNs) demonstrated hugely improved, state-of-the-art accuracy when applied to image classification. Following this discovery, a large portion of the entrants and all of the subsequent winners have been based on deep learning with CNNs. Sermanet et al. [3] presented a system named OverFeat in 2013 to classify images whilst simultaneously providing localisation information and object detection, adding a large amount of information to the output of classification attempts. The proposed system used a CNN combined with multiscale sliding window processing to achieve this. Girshick et al [4] created a system named Regions with CNN features (R-CNN) in 2014 which also uses region specific object classification rather than image-wide classification. The R-CNN uses images alongside accompanying proposed regions (as locations in the image) as inputs. The images are then cropped, creating individual sub-images for the proposed regions. The sub-images can then be fed through a CNN for processing. Classification is performed in the final fully connected layer using a Support Vector Machine (SVM) for each object. R-CNN provides a significant improvement in error rates when compared with OverFeat, due to the use of selective search region proposal techniques, rather than the sliding window approach used by OverFeat. The use of a region-specific object classifier allows for much finer-grained description generation through the inclusion of location data which can provide relative position information for any objects or people retrieved from the image. Girshick [5] recently improved the original R-CNN to create Fast R-CNN, with a number of changes resulting in greatly improved classification performance.

3 The Proposed Intelligent Chat

The proposed system is designed to function as a web application in order to reduce potential barriers to use due to the widespread ownership of Internet accessible, camera equipped devices. The functionality is separated over the client-server architecture in an attempt to make the most of available processing power on both sides, allowing for responsive communication. The application was designed to also function as an information retrieval system, enabling users to ask questions which could be answered using the Internet as a knowledgebase. The accessibility of computer systems, whilst constantly improving, can still be an issue which prevents their use by the elderly. By incorporating access to the Internet and to wider functionality, through our accessible audial and oral interface, we provide a much less daunting way to assimilate these new users into the connected world.

3.1 The System Architecture

The system was designed as a distributed model with a view to operate from a central server, allowing access to the functionality by the end users through a web browser.

The server itself is a Python application built upon Tornado [6], an asynchronous, real-time, web framework. Communication between the central server and the individual client web browsers is performed through the use of websockets. On the server end, Tornado provides access to the websocket functionality, allowing interaction through Python methods. On the client-side, this interaction is provided by the Javascript library Socket.io. The image capture functionality is implemented using the getUserMedia method in HTML5 to capture webcam data. All of the image capturing and streaming are performed on the client-side using HTML5 and Javascript. Overall, the agent-based system incorporates the server-based image description generation and Wikipedia question answering alongside the client-based emotional expression recognition.

3.2 Conversation Extensions

The conversational functionality forms the core of the system, providing the avatar with a means of prolonging interaction with the user and attempting to maintain a flow of conversation. The conversational system is implemented using Artificial Intelligence Markup Language (AIML) [7]. AIML provides an XML compliant framework, allowing for the creation of complex two-way chat functionality without scripting every response, achieved through the use of recursive pattern matching [8]; resulting in a realistic, albeit simple way of creating a conversational system. AIML provides a number of open source libraries containing pre-written chat functionality and conversation which can be extended to suit the particular use case. The most prominent of these is known as 'The Artificial Linguistic Internet Computer Entity', i.e. A.L.I.C.E. We used the ALICE library as a starting point for the companion's chat functionality and extended its conversational capabilities by incorporating our own AIML.

AIML provides the tools to set up artificial two-way communication using text. With our goal of creating an approachable companion to be used by elderly, not necessarily computer literate users, we decided to bridge the accessibility gap by providing audial and oral communication via speech recognition and synthesis. Both aspects were implemented using the relatively new HTML5 Web Speech API [9]; audial using the SpeechSynthesis interface while oral using the SpeechRecognition interface. Interaction with the system can therefore be performed entirely through spoken conversation.

We have extended ALICE's vocabulary through the implementation of a question-answering system, using Wikipedia as a data source. The system currently parses questions and searches Wikipedia using the main terms of the question if it cannot answer using its existing vocabulary. A short summary of the search term from the beginning of the Wikipedia page is then retrieved and spoken to the user.

The proposed system also includes functionality to retrieve and present the user's location using the HTML5 geolocation API. The location is spoken using latitude and longitude co-ordinates – a small map image is also illustrated showing the user's location.

3.3 Image Description Generation

The proposed Intelligent Chat system has included a novel image description generation component. It is implemented using a deep learning architecture to provide natural language description of the content of an image. It includes a number of key steps; object detection and recognition, attribute prediction, scene classification, and description generation. The system architecture of this image description generation component is provided in Fig. 2. We discuss each key step of this image description generation function in detail below.

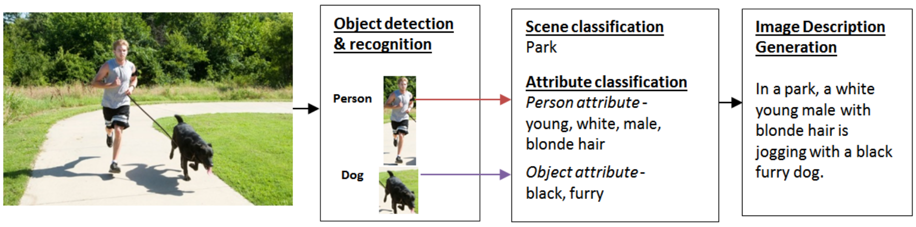

Object detection & recognition
Person
Dog
Scene classification
Park
Attribute classification
Person attribute - young, white, male, blonde hair
Object attribute - black, furry
Image Description Generation
In a park, a white young male with blonde hair is jogging with a black furry dog.

Fig. 2. The architecture for deep learning based image description generation

Object Detection & Recognition. The first stage of the proposed image description generation component is to conduct object detection and recognition. The object detector implemented is the Regional Convolutional Neural Network, i.e. R-CNN, from Girschick et al. [4] consisting of 8 learned layers; 5 convolutional layers and 3 fully connected layers. The output of this object detection function is to identify bounding box coordinates of each salient object in the image. This network can detect and classify 200 object categories from the ImageNet 2013 dataset, collecting selective search data from the whole image. These regions are then each classified by 200 SVMs in order to determine which areas contain a specific object in an image. The object detector is pre-trained on the ILSVRC-13 object detection challenge, taking approximately a week to train on state-of-the-art hardware.

Attribute Prediction. Creating a full sentence description of the visual image input of users' environment is a more challenging task than object labelling. To achieve this, the object image(s) detected from the previous stage are passed to another CNN provided by Chatfield et al. [10]. This network has a similar structure to normal CNNs, however does not possess the fully connected or classification layers at the lower end of the network, meaning that the network only extracts image feature vectors to be used elsewhere. Like the R-CNN, this network consists of 8 learned layers; 5 convolutional layers and 3 fully connected layers.

The collected features are then used to train multiple attribute classifiers in order to increase the descriptiveness of an object label. There are more than 50 attributes used in this research in total, in order to provide more detailed people and object descriptions. For objects, there are 26 attributes relating to colour, shape and size information. These attributes are collected from a fully annotated subset of the ImageNet dataset [11]. There are also 26 attributes for human description ranging from hair colour, style, age and ethnicity. These are taken from the PubFig dataset [12], originally consisting of more than 70 attributes to describe people. For these experiments only a subset of these attributes (i.e. 26 attributes) are used during testing for description generation.

Scene Classification. Scene recognition and classification is used to enhance the image description generation by providing an overall idea of the setting, which can then be enhanced through the inclusion of detailed object and person description. The classification system used is a CNN trained on the MIT Places dataset created by Zhou et al [13]. The Places dataset contains over 7 million images from 476 scene categories created using Amazon Mechanical Turk workers to label the images. The network proposed by Zhou et al. has been used in this research, which was trained on almost 2.5 million images, comprised of 205 different scene categories. Its architecture was originally taken from the Caffe reference network [14].

Facial Expression Recognition. The system has integrated an intelligent facial expression recognition component to identify seven basic emotions: happiness, anger, sadness, disgust, surprise, fear, and contempt. It borrows the architecture implemented in [15, 16, 17] and consists of feature extraction using Local Binary Patterns and micro-GA embedded Particle Swarm Optimization feature selection. It has been proven to show superior performances in comparison to related research when evaluated with the CK+ dataset [18].

Sentence Generation. To construct a valid descriptive sentence, the recognized object and attribute labels must be combined in a very natural-sounding manner. In this work, a template-based approach is used to transform these descriptive labels into multiple short sentences that can be concatenated and reported as a single detailed description of the image in question. The scene label collected in the earlier stage is also utilised in this process, which is used either as an opening or a closing statement to the sentence. An example output of the deep-learning based image description generation is shown in Fig. 2.

Fall and Hazard Detection. The image description framework also has the capability to describe out-of-scope images such as hazardous objects on the floor and a falling person. Both of these aspects are based on the assumption that the camera is at a fixed height and a threshold value has previously been determined where the floor meets the wall. A hazardous object and fall actions are reported when a detected object or person is reported below the threshold value. These are again described as sentences and merged with the previous outputs to ask users if they require assistance, or to suggest that the hazards should be moved to safer locations. An example for fall action description is provided in Fig. 3.

4 Evaluation

A lightweight version of the system without image description generation was used for evaluation in order to compare its performance with that of the full version of the system with image description generation during user evaluation. This cut down version of the system was hosted on an AWS t2 micro instance to allow for distributed testing and uptime evaluation. The system successfully handles multiple simultaneous users from geographically diverse locations. No system crashes were observed over the course of several months.

The overall system with image description generation has been hosted online on a server possessing a relatively powerful GPU to enable full access to the entire functionality of the system. Client Internet access is also required for the speech synthesis and recognition portions of the client-side functionality. The full system also successfully handles multiple concurrent users from geographically diverse locations. The system has been successfully intensively tested under a vast number of real-life settings. Evaluation results for image description generation using the proposed deep learning architecture are discussed in detail below.

4.1 Evaluation of Image Description Generation

The ROUGE score [19] (Recall-Oriented Understudy for Gisting Evaluation) is used to evaluate the image description generation component of the proposed system. It provides a metric to determine the quality and similarity of summaries between human description annotations and computer generated outputs. The metrics within ROUGE are based on the number of n-grams, sequences of words and word pairs between machine-generated and ground truth summaries.

In order to evaluate the proposed system, an existing image description dataset is used, i.e. the IAPR-12 dataset [20], which consists of ~20,000 images, each annotated with a descriptive sentence or description in both English and German. Specifically, a subset of 100 images from IAPR-12 is used for the evaluation of our work to generate the ROUGE score.

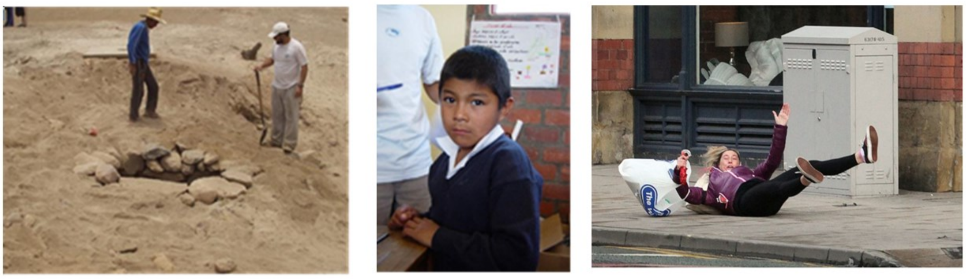
ROUGE-1 and ROUGE-2 are n-gram based approaches, essentially favouring generated descriptions that contain n-grams shared with the Ground Truth (GT) descriptions, thus preferring a description which is similar to the GT sentences. ROUGE-L refers to the Longest Common Subsequence (LCS), which considers two sequences X and Y. This LCS, is the subsequence that occurs in both sequences with the maximum length. In sentence level ROUGE-L, the perception is that the longer the LCS of generated and GT sentences, the more similar the sentences. The ROUGE scores shown in Table 1, show similar and in some cases improved scores over existing image description generation systems, such as [21], that also implement this metric. Fig. 3 shows some example outputs of the proposed image description generation component.

Since the proposed image description generation function has a series of individual key steps with each trained on datasets relevant to their intended use but unrelated to the whole images, it is able to detect and recognize a large number of object classes especially from challenging cross-domain images, ensuring a minimal amount of data loss. Pairing this with the ability to recognise and classify a large number of human and object attributes enables the system to create descriptive attribute labels and produce descriptive sentences, of any image the system processes. Moreover, other existing image description generation systems tend to be trained and tested on the same or very similar datasets [22, 23], meaning the methods used in these frameworks essentially understand the types of sentences and structures required in order to achieve higher scores for test images from the same domain, however testing these frameworks with irrelevant cross-domain images tends to produce unsatisfactory results with dramatically reduced performances. In comparison with the above related research, the proposed system focuses on a regional approach and experimental results indicate that it shows great robustness and flexibility in dealing with out-of-scope or cross-domain image description generation tasks because of its focusing on image regional details to retrieve more local information. Therefore, the scores achieved by the Intelligent Chat system are significant, especially when dealing with cross-domain (e.g. healthcare) images.

Table 1. ROUGE scores over the 100 images by comparing system generated results to GT descriptions provided by the IAPR-12 image database. For each metric, the recall (R), precision (P) and F-scores (F) averaged over the 100 images are presented.

	ROUGE-1			*ROUGE-2*			*ROUGE-L*		
100 Images	R	P	F	R	P	F	R	P	F
Results	0.220	**0.389**	0.281	0.041	0.072	0.052	0.210	**0.372**	0.269

Another preliminary user evaluation with 50 users is also conducted using the two versions of the system with and without image description generation. The version with image description generation achieved higher user satisfaction and significantly improved user experience than the version without image analysis. Moreover, the overall system with image description generation resulted in positive comments regarding the integration of Wikipedia and image description generation to enhance question answering and human agent interaction. Most users agreed that the image description generation function would be useful for visually impaired users and the detection of falls and hazards would be especially helpful in assisting independent living.



In a desert, I can see a middle-aged white male with blonde hair is next to a senior white male with grey hair.

In a kindergarten classroom, I can see a child white male with black hair

In a plaza, there is a young white male with brown hair. I can see a person on the floor. Do you require assistance?

Fig. 3. Example image descriptions generated by the proposed system

5 Conclusion and Future Work

This research proposes a vision enriched Intelligent Chat system for elderly care. The system is developed to conduct facial emotion recognition, object and scene recognition, hazardous objects and scene classification, and fall detection. Deep learning based image description generation is also used to generate sentences based on the above outputs to warn of hazards or generate alarms when falls occur. The Intelligent Chat system is tested with users in real-life settings and evaluation results indicate that it achieves 0.389 ROUGE-1 score for image description generation for the evaluation of 100 images from the IAPR-12 dataset, which is comparable to other state-of-the art related research [21]. Future implementations of this work could be improved by utilising a faster, more efficient object detector such as the Fast R-CNN. The sentence generation functionality could also be altered to use machine learned methods such as Recurrent Neural Networks that have proven successful in machine translation [24, 25].

The proposed system could be further extended to perform the observational duties of a carer, allowing a single carer to be alerted to accidents or incidents involving any of a number of distinct patients, in potentially separate geographical locations. Such a system could enable a greater degree of independence for patients who would otherwise require constant human supervision. Alternatively, the techniques applied in this work could be applied in other domains in order to enhance human-computer interaction, such as workplace training or interactive learning for children. Moreover, the proposed health monitoring system could be extended to integrate with diagnostic systems where images (e.g. retinal or blood images) can be taken by smart devices for analysis to promote early diagnosis [26, 27].

Acknowledgments. We appreciate the funding support received from Higher Education Innovation Fund and RPPTV Ltd.

References

1. De Vos, E. 2002. Look at that doggy in my windows, on effects of anthropomorphism in human-agent interaction, Doctoral Thesis, Utrecht University.
2. Krizhevsky,A., Sutskever, I. and Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. 1097-1105. (2012).
3. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. and LeCun, Y.: OverFeat: Integrated recognition, localization and detection using convolutional networks. (2013).
4. Girshick, R., Donahue, J., Darrell, T. and Malik, J.: Rich feature heirarchies for accurate object detection and semantic segmentation. IEEE Transactions on Computer Vision and Pattern Recognition (CVPR), 580-587. (2014).
5. Girshick, R.: Fast R-CNN. In Proceedings of ICCV 2015. (2015).
6. Facebook.: Tornado. Available at:http://www.tornadoweb.org/en/stable/, (2011).
7. Wallace, R.: The elements of AIML style. Alice AI Foundation. Available at: https://files.ifi.uzh.ch/cl/hess/classes/seminare/chatbots/style.pdf, (2003).
8. Wallace, R.: Symbolic reductions in AIML. Available at: http://www.alicebot.org/documentation/srai.html (2000).

9. Shires, G. and Wennborg, H.: Web speech API specification. W3C Community Final Specification Agreement. https://dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html, (2012).

10. Chatfield, K., Simonyan, K., Vedalsi, A. and Zisserman, A.: Return of the devil in the details delving deep into convolutional neural nets. BMVC (2014).

11. Fei Fei, L.: ImageNet: crowdsourcing, benchmarking & other cool things. CMU VASC Seminar. (2010).

12. Kumar, N., Berg, A.C., Belhumeur, P.N. and Nayar, S.K.: Attribute and simile classifiers for face verification. International Conference on Computer Vision (ICCV). (2009).

13. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A. and Oliva, A.: Learning deep features for scene recognition using places database. Advances in Neural Information Processing Systems, 487-495. (2014).

14. Jia, Y. et al.: Caffe: An open source convolutional architecture for fast feature embedding. Available at: http://caffe.berkeleyvision.org/, (2013).

15. Neoh, S.C., Zhang, L., Mistry, K., Hossain, M.A., Lim, C.P., Aslam, N. and Kinghorn, P.: Intelligent facial emotion recognition using a layered encoding cascade optimization model. Applied Soft Computing, 34(2015), 72-93. (2015).

16. Mistry, K., Zhang, L., Neoh, S.C., Lim, C.P. and Fielding, B.: A micro-GA Embedded PSO Feature Selection Approach to Intelligent Facial Emotion Recognition. IEEE Transactions on Cybernetics. ISSN 2168-2267. 1-14. (2016)

17. Zhang, L., Mistry, K., Jiang, M., Neoh, S.C. and Hossain, A.: Adaptive facial point detection and emotion recognition for a humanoid robot. Computer Vision and Image Understanding, 140. 93-114. (2015).

18. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z. and Matthews, I.: The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. In Proceedings of CVPR4HB. (2010).

19. Lin, C.: ROUGE: A package for automatic evaluation of summaries. In Proceedings of Workshop on Text Summarization Branches Out,. (2004).

20. Grubinger, M., Clough, P.D., Müller, H. and Deselaers, T.: The IAPR Benchmark: A new evaluation resource for visual information systems. International Conference on Language Resources and Evaluation. (2006).

21. Lin, D., Fidler, S., Kong, C. and Urtasun, R.: Generating multi-sentence natural language descriptions of indoor scenes. British Machine Vision Conference (BMVC). (2015).

22. Karpathy, A. and Fei Fei, L.: Deep visual-semantic alignments for generating image descriptions. Computer Vision and Pattern Recognition (CVPR). (2015).

23. Vinyals, O., Toshev, A., Bengio, S. and Erhan, D.: Show and Tell: A neural image caption generator. Computer Vision and Pattern Recognition (CVPR). (2015).

24. Sutskever, I., Vinyals, O. and Le, Q.V.: Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems. 3104-3112. (2014).

25. Jozefowicz, R., Zaremba, W. and Sutskever, I.: An Empirical Exploration of Recurrent Network Architectures. In Proceedings of the 32nd International Conference on Machine Learning (ICML-15), 2342-2350. (2015).

26. Neoh, S.C., Srisukkham, W., Zhang, L, Todryk, S., Greystoke, B., Lim, C.P., Hossain, A. and Aslam, N.: An Intelligent Decision Support System for Leukaemia Diagnosis using Microscopic Blood Images. Scientific Reports, 5 (14938). (2015).

27. Bourouis, A., Feham, M., Hossain, M. A. and Zhang, L.: An Intelligent Mobile based Decision Support System for Retinal Disease Diagnosis. Decision Support Systems. 59, 341–350. (2014).