A Vision Enriched Intelligent Agent with Image Description Generation (Demonstration)

Li Zhang, Ben Fielding, Philip Kinghorn and Kamlesh Mistry
Department of Computer Science and Digital Technologies
Faculty of Engineering and Environment
Northumbria University
Newcastle, UK, NE1 8ST
{li.zhang; ben.fielding; philip.kinghorn; kamlesh.mistry}@northumbria.ac.uk

ABSTRACT
In this paper, we present an intelligent conversational agent enriched with automatic image understanding and facial expression recognition using 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 capable of conducting question-answering via speech-based interaction, but also able to provide analysis of the user's surroundings, emotional states, hazards and fall actions via visual data. The agent is accessible from a web browser and can be communicated with via voice or text 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
Human-agent Interaction; Image Understanding; Agent Architectures

1. INTRODUCTION
We propose an intelligent emotion and health well-being monitoring system for elderly care by incorporating a number of computer vision techniques including scene and object recognition, image description generation, and emotion recognition, with an approachable intelligent conversational 3D humanoid avatar.

The proposed system known as 'Intelligent Chat' provides real-time visual, audial and oral conversation with a number of additional features tailored to the needs of elderly users. It analyzes users' emotional states using an intelligent facial emotion recognition component applied to the user's webcam, integrated in the browser. Moreover the user's environment is analyzed by the proposed system using more advanced vision-based deep learning techniques embedded in a central server, 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 to warn hazards and danger, thereby providing an engaging and life-like companionship experience. Intelligent Chat also provides answers to queries on any subject using integration with the popular online encyclopedia Wikipedia, enabling conversation to cover potentially any topic. Empirical results indicate that the proposed system compares favourably with other state-of-the-art related applications in terms of image description generation, facial expression recognition and intelligent health monitoring applications.

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.),* May 9–13, 2016, Singapore.

2. RELATED WORK

2.1 Image Description Generation
Image description generation, or the component parts of such a system, has been the focus of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) since its conception in 2010. Krizhevsky et al. [1] indicated that deep Convolutional Neural Networks (CNNs) could demonstrate hugely improved, state-of-the-art accuracy when applied to image classification. Following this research finding, a large portion of the entrants and many subsequent state-of-the-art developments have been based on such deep learning architectures with CNNs as encoders. Moreover, Sermanet et al. [2] proposed OverFeat to classify images whilst simultaneously providing localization information and object detection, adding a large amount of information to the output of classification attempts. Their system integrated a CNN with multiscale sliding window processing for object detection. Moreover, there are also other developments in the field using encoder and decoder deep learning architectures for solving diverse complex computer vision tasks [3, 4].

2.2 Facial and Bodily Expression Recognition
Many state-of-the-art research applications have been proposed recently for facial and bodily expression recognition. Neoh et al. [5] developed a facial expression recognition system with direct similarity and Pareto based optimization whereas Zhang et al. [6] developed a novel facial point detector based on unsupervised

learning for intensity estimation of facial action units for diverse challenging cases, such as occlusions, rotations and illumination changes. Zhang et al. [7] also proposed a bodily expression recognition system embedded with a particle swarm optimization variant to conduct regression of valence and arousal dimensions.

In comparison to the above state-of-the-art related work, this research employs a novel regional deep learning structure for image description generation to improve existing image captioning techniques, by including regional object detection and recognition, scene classification, and template-based sentence generation. It also incorporates both texture and shape information for facial expression recognition with the attempt to address limitations of single modality based methods. The experimental results proved the superiority of our approach for both image description generation and facial expression recognition.

3. THE PROPOSED SYSTEM

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 object and scene classification, and fall detection. Template-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.

We design the system to function as a web-based application with 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. The application is designed to also function as an information retrieval system, enabling users to ask questions which could be answered using the internet as a knowledge base, allowing for limitless knowledge to be accessed through the chat to enhance user experience. The interface of the proposed 'Intelligent Chat' system is shown in Figure 1.

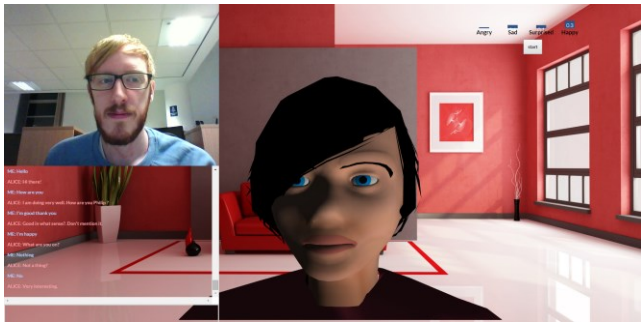

Figure 1. The interface of the 'Intelligent Chat' system

3.1 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, rather than simply reacting to questions. The conversational aspect is implemented using Artificial Intelligence Markup Language (AIML) [8, 9, 10]. Related APIs have also been implemented to enable speech synthesis and recognition interface developments.

The AIML vocabulary, whilst containing a large amount of information and responses, is noticeably incomplete, due to the restrictions of the AIML language. We have extended the agent's vocabulary monumentally through the implementation of a question-answering system using Wikipedia as a data source. This allows the intelligent agent to answer questions on practically any topic available through collective human knowledge. We also include a new functionality to retrieve and present the user's location using the HTML5 geolocation API. The location is spoken using latitude and longitude co-ordinates with a small map image embedded in the interface to indicate the user's location.

3.2 Image Description Generation

3.2.1 Object Detection & Recognition

The first stage of the image description generation embedded in Intelligent Chat is based upon object detection. The object detector implemented is the Regional Convolutional Neural Network, i.e. R-CNN, from Girschick et al. [11], consisting of 8 learned layers, 5 convolutional layers and 3 fully connected layers. This network is further extended to detect and classify 200 object categories from 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. This object recognition process is also extended to conduct hazardous objects and fall action detection.

3.2.2 Scene Recognition

Scene classification is used to enhance the image description generation by providing an overall context 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 [3]. The Places dataset contains over 7 million images from 476 scene categories. The network proposed by Zhou et al [3] was used in our work, which was trained on 2.5 million images, comprised of 205 scene categories.

3.2.3 Facial Emotion Recognition

A facial expression recognition component is also integrated into Intelligent Chat. Active Appearance Model and Neural Networks are used for the recognition of 7 emotions including; happiness, anger, sadness, disgust, surprise, fear and contempt. 3000 images extracted from CK+ are used for the training of AAM [12, 13].

3.2.4 Sentence Construction

To construct a valid descriptive sentence, the recognized object labels, scene context and emotions must be combined in a very natural-sounding manner. In this work, a template-based approach using rule-based reasoning 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 context identified is also used either as an opening or a closing statement of the generated sentence.

We have tested Intelligent Chat with both real users in real-life settings and hundreds of real-world images. Evaluation results indicate that it achieves 0.389 ROUGE score for image description generation which is impressive and comparable to other state-of-the-art research [14]. Future implementations of this work could be improved by utilising a faster, more efficient object detector such as the Fast R-CNN, and pairing it with a more efficient sentence generation method such as Recurrent Neural Networks [15]. Moreover, the system could be further extended to integrate with disease diagnosis functions [16, 17]. We also aim to conduct substantial system evaluation with elderly users to further prove system efficiency. Demo URL:

https://drive.google.com/file/d/0BxS6ufv4G_TXYmxUSVRiTS1tUUk/view?usp=sharing

4. REFERENCES

[1] Krizhevsky, A., Sutskever, I. and Hinton, G.E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*. 1097-1105.

[2] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. and LeCun, Y. 2014. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *International Conference on Learning Representations*.

[3] Zhou, B., Lapedriza, A., Xiao, J., et al. 2014. Learning Deep Features for Scene Recognition using Places Database. *Advances in Neural Information Processing Systems*.

[4] Jia, Y. et al. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *Proceedings of MM'14*, 675-678.

[5] Neoh, S.C., Zhang, L., Mistry, K., Hossain, M.A., Lim, C.P., Aslam, N. and Kinghorn, P. 2015. Intelligent Facial Emotion Recognition Using a Layered Encoding Cascade Optimization Model. *Applied Soft Computing*, 34, 72-93.

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

[7] Zhang, Y., Zhang, L., Neoh, S.C., Mistry, K. and Hossain, A. (2015). Intelligent affect regression for bodily expressions using hybrid particle swarm optimization and adaptive ensembles. *Expert Systems with Applications*, 42. 8678-97.

[8] Wallace, R. 2003. The elements of AIML style.

[9] Zhang, L. and Barnden, J. (2012). Affect Sensing Using Linguistic, Semantic and Cognitive Cues in Multi-threaded Improvisational Dialogue. *Cognitive Computation*. Volume 4, Issue 4, 436-459.

[10] Zhang, L. and Barnden, J.A. (2010). Affect and Metaphor Sensing in Virtual Drama. *International Journal of Computer Games Technology*. Vol. 2010. Article ID 512563.

[11] Girshick, R., Donahue, J., Darrell, T. and Malik, J. 2014. Rich Feature Heirarchies for Accurate Object Detection and Semantic Segmentation. *CVPR*, 580-587.

[12] Zhang, Y., Zhang, L. and Hossain, M.A. (2015). Adaptive 3D facial action intensity estimation and emotion recognition, *Expert Systems with Applications*, 42 (2015).

[13] Zhang, L., Jiang, M., Farid, D. and Hossain, A.M. (2013). Intelligent Facial Emotion Recognition and Semantic-based Topic Detection for a Humanoid Robot. *Expert Systems with Applications*, 40 (2013), 5160-5168.

[14] Lin, D., Kong, C., Fidler, S. and Urtasun, R. (2015). Generating multi-sentence lingual descriptions of indoor scenes. In *British Machine Vision Conference*. UK.

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

[16] Bourouis, A., Feham, M., Hossain, M. A. and Zhang, L. (2014). An Intelligent Mobile based Decision Support System for Retinal Disease Diagnosis. *Decision Support Systems*. Elsevier. Volume 59, March 2014, 341–350.

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