# Multimodal Representation Learning for Human Robot Interaction

Eli Sheppard
ems7@hw.ac.uk
Edinburgh Centre for Robotics
Edinburgh, UK

Katrin. S. Lohan
k.lohan@hw.ac.uk

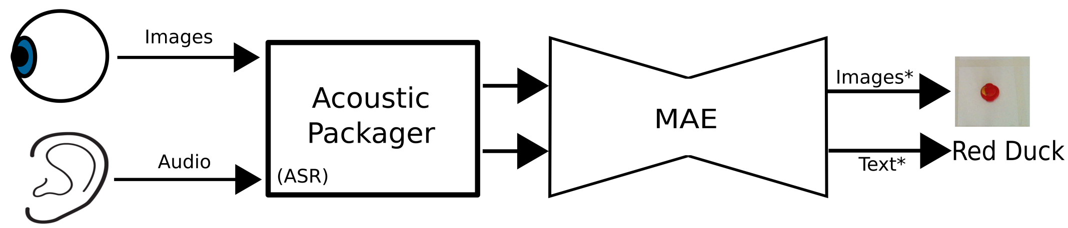


**Figure 1: System schematic. Data is captured from sensors by an acoustic packager and fed to the multimodal autoencoder (MAE).**

## ABSTRACT

We present a neural network based system capable of learning a multimodal representation of images and words. This representation allows for bidirectional grounding of the meaning of words and the visual attributes that they represent, such as colour, size and object name. We also present a new dataset captured specifically for this task.

## CCS CONCEPTS

• **Computing methodologies → Vision for robotics**; **Neural networks**; *Natural language processing*; *Cognitive robotics*.

## KEYWORDS

datasets, neural networks, unsupervised learning, symbol grounding, robotics

**ACM Reference Format:**
Eli Sheppard and Katrin. S. Lohan. 2020. Multimodal Representation Learning for Human Robot Interaction. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20 Companion), March 23–26, 2020, Cambridge, United Kingdom.* HRI 2020, Cambridge, UK, 2 pages. https://doi.org/10.1145/3371382.3378265

## 1 INTRODUCTION

In order for robots to become ubiquitous, they must be able to cope with learning to identify new objects continuously without human intervention. We present a novel method capable of learning a joint representation across the visual and textual modalities which can be exploited to allow robots to learn the visual attributes of objects and the words used to describe them in a grounded manner. [1, 2]. This is known as Multimodal Representation Learning (MRL) [9].

We provide a new dataset called Real-Shapes (ReShape) which contains 7 objects, in 10 colours and 3 sizes. Not all objects appear in all 10 colours or all 3 sizes.

## 2 METHOD

### 2.1 Data Acquisition

The Real-Shapes dataset (ReShape) was created by presenting various objects to a webcam in 9 different locations and giving a short, verbal description of the object [1].

Data is captured using a webcam and microphone. The data is packaged together using Acoustic Packaging [7, 8]. Speech captured by the microphone at 16kHz is transcribed using Automatic Speech Recognition (ASR). Each transcribed utterance contains the size, colour, name and location of the object presented to the webcam.

Images are captured at 10 frames per utterance, 640x480 pixels and then cropped to 200x200 pixels, based on the uttered location so that the object is roughly centred in the crop. Cropped images are then rescaled to 64x64 pixels and locations are removed from the utterances such that each utterance is of the form <size> <colour> <name>. Transcribed utterances are then encoded as binary vectors with 1 representing the presence of a word in the description. The MAE has a 20 word vocabulary.

### 2.2 Training Procedure

To learn a grounded multimodal representation a subset of the ReShape data is used to train a Multimodal Autoencoder (MAE) [6, 9, 10]. The MAE consists of stacked layers of convolution, batch normalisation and dropout [11], with two inputs and two outputs (one each for images and text).

Pairs of images and their descriptions are fed to the MAE and their embeddings are merged by concatenation, after several layers of convolution [2]. After merging the two modalities, two decoder


*HRI '20 Companion, March 23–26, 2020, Cambridge, United Kingdom*

ACM ISBN 978-1-4503-7057-8/20/03.
https://doi.org/10.1145/3371382.3378265

[1]The dataset can be downloaded from https://bit.ly/38lNh37
[2]A full implementation can be found at https://bit.ly/341fBo8

branches work to reproduce the original image and text inputs as seen in Figure 2.

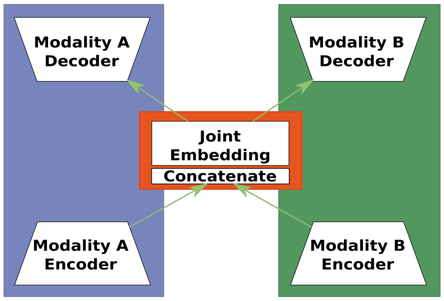


**Figure 2: A Multimodal Autoencoder.**

Data is provided to the MAE in three ways Bimodal (Bi), Image Only (Im) and Words Only (Wo). The MAE is trained to generate image and text outputs regardless of whether both images and text are provided as input (Bi), only images are provided as input (Im) or only text is provided as input (Wo). Data is provided in all three manners during training, essentially tripling the number of training examples.

To improve the quality of the generated images, target images are replaced with class exemplars when only words are provided as input. Exemplars are selected by calculating the mean image for each object-colour-size combination from the training data and selecting the image closest to the mean.

## 3 RESULTS

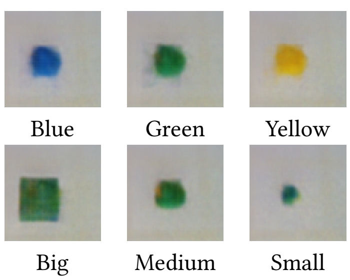


**Figure 3: Images generated from individual words.**

Figure 3 shows images generated by the MAE from individual words. The MAE has correctly learnt the meanings of these words; given the word "Blue" it generates blue pixels, "Green", it generates green pixels and "Yellow", yellow pixels. Further to this, we see that given the word "Big" it generates lots of coloured pixels, "Medium", less coloured pixels and "Small", the least coloured pixels.

The MAE also correctly learns the meanings of the names of the different objects and combinations of colours, sizes and object names, even ones unseen in the training data (Figure 4).

## 4 CONCLUSION AND FUTURE WORK

We present a novel system capable of learning the grounded meaning of different visual attributes (Size, Colour, Shape) and their textual equivalents. In this preliminary experiment we show how


**Figure 4: Different sized donuts that don't appear in the training data.**

this method can generalise to unseen combinations of colours, sizes and shapes.

The performance of the MAE on the test data will be evaluated in future work.

In future work we will utilise the system in an interactive scenario using the iCub robot. To do this, we have implemented a Natural Language Understanding (NLU) system which allows humans to query the MAE through conversation with the robot about the colour, size and name of different objects as well as to interactively teach the iCub new objects.

Switching to a Word2Vec [5] encoding of language instead of the binary one used here will allow for an expanding vocabulary.

We will also continue to collect data for the dataset, covering more diverse lighting conditions, different backgrounds and more objects in order to enhance the quality of the multimodal embedding learnt by the MAE [3, 4].

## 5 ACKNOWLEDGMENTS

This work was funded by the EPSRC.

## REFERENCES

[1] Broz, F., Nehaniv, C. L., Belpaeme, T., Bisio, A., Dautenhahn, K., Fadiga, L., Ferrauto, T., Fischer, K., Förster, F., Gigliotta, O., et al. The italk project: A developmental robotics approach to the study of individual, social, and linguistic learning. *Topics in cognitive science 6*, 3 (2014), 534–544.

[2] Cangelosi, A., Belpaeme, T., Sandini, G., Metta, G., Fadiga, L., Sagerer, G., Rohlfing, K., Wrede, B., Nolfi, S., Parisi, D., et al. The italk project: Integration and transfer of action and language knowledge in robots. In *Proceedings of Third ACM/IEEE International Conference on Human Robot Interaction (HRI 2008)* (2008), vol. 12, p. 15.

[3] Keller, I., and Lohan, K. S. Analysis of illumination robustness in long-term object learning. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (2016), IEEE, pp. 240–245.

[4] Keller, I., and Lohan, K. S. On the Illumination Influence for Object Learning on Robot Companions. *Frontiers in Robotics and AI,(in press)* (2019), 1–17.

[5] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (2013), pp. 3111–3119.

[6] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (2011), pp. 689–696.

[7] Schillingmann, L., Wrede, B., and Rohlfing, K. Towards a computational model of acoustic packaging. In *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on* (2009), IEEE, pp. 1–6.

[8] Schillingmann, L., Wrede, B., and Rohlfing, K. J. A computational model of acoustic packaging. *IEEE Transactions on Autonomous Mental Development 1*, 4 (2009), 226–237.

[9] Sheppard, E., Lehmann, H., Rajendran, G., McKenna, P. E., Lemon, O., and Lohan, K. S. Towards life long learning: Multimodal learning of mnist handwritten digits. *IEEE ICDL EPIROB 2018 Workshop on Life Long Learning* (2018).

[10] Silberer, C., and Lapata, M. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2014), vol. 1, pp. 721–732.

[11] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research 15*, 1 (2014), 1929–1958.