Multimodal Learning: Examples in Gesture and Audio-Visual Speech Recognition

Hsieh Yu-Guan

August 7, 2017

Abstract

1 Introduction

2 Related Work

3 Presentation of Basic Network Architectures

4 Datasets and Preprocessing

4.1 Creative Senz3D

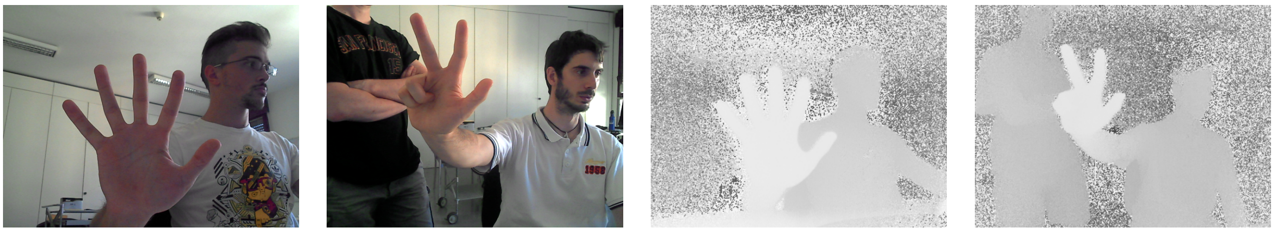

Figure 1: **Example images in the Creative Senz3D dataset.**
Left Two) Color images.
Right Two) Corresponding depth images.
All of the images are of size 480×640 and contain the the entire upper body of the subject.

4.2 ASL Finger Spelling

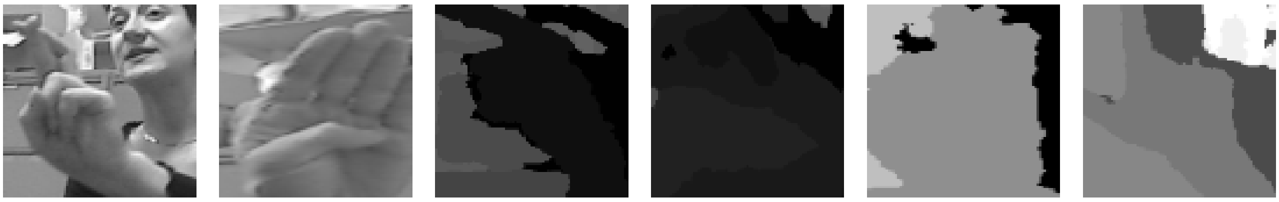

Figure 2: **Example images in the ASL Finger Spelling dataset (after preprocessing).**
Left Two) Grayscale intensity images.
Middle Two) Depth maps after adjusting contrast.
Right Two) Depth maps after Z-normalization.
Images of this dataset have variable sizes, and they're all resized to 83×83 before being fed to the network. Generally only the hand region is contained in image.

4.3 AVletters

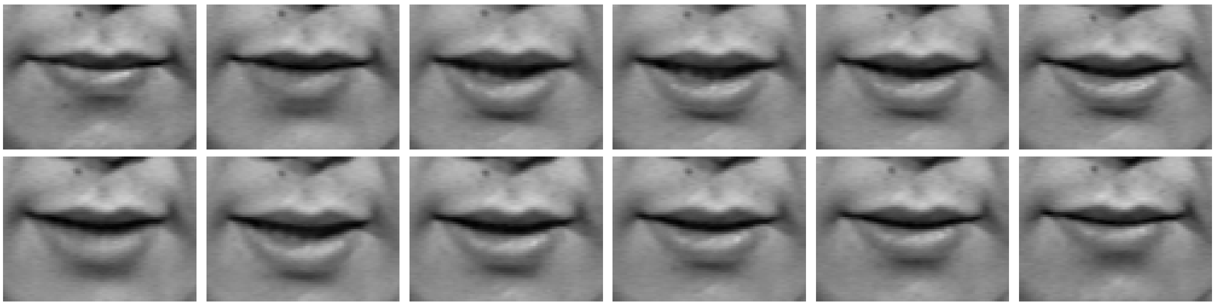

Figure 3: **Example visual input for the AVletters dataset (left to right, top to bottom).** Pre-extracted lip regions of 60×80 pixels are provided. Each image sequence is resampled to be of length twelve in order to give an input of fixed size to the network.

5 Experimental Setup

6 Experiences and Results: Unimodal Cases

6.1 Classification

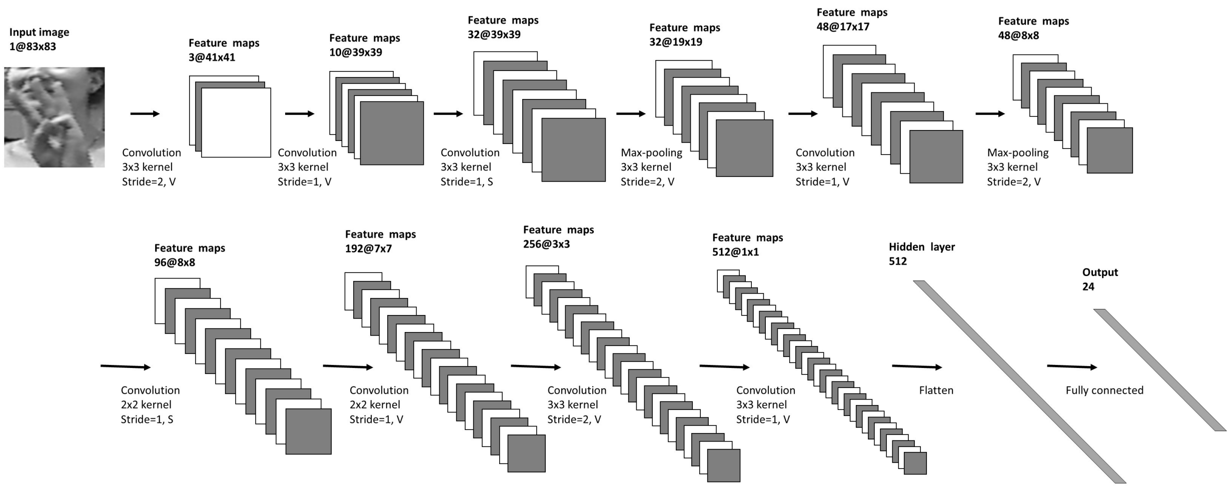

Figure 4: **CNN architecture used for the Finger Spelling dataset.**
The input of the nework is a one-channel image of size 83×83. It contains ten hidden layers. S stands for 'SAME' padding and V stands for 'VALID' padding (see text).

6.2 Convolutional auto-encoder

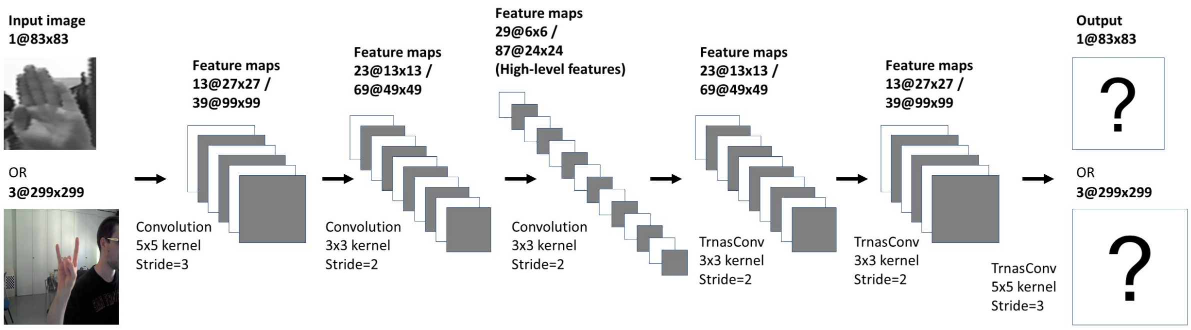

Figure 5: **Convolutional auto-encoder architecture with three convolutional layers and three tranposed convolutional layer.**
Activation values of the middle layer are taken as high-level features of the input image. Inputs of the network can be of different sizes. We only use valid paddings here.

Figure 6: **Image restoration using convolutional auto-encoder.**
Left) Clean Image.
Middle) Noisy image [input].
Right) Restored image [output].

7 Experiences and Results: Multimodal Cases

7.1 Learning shared representation

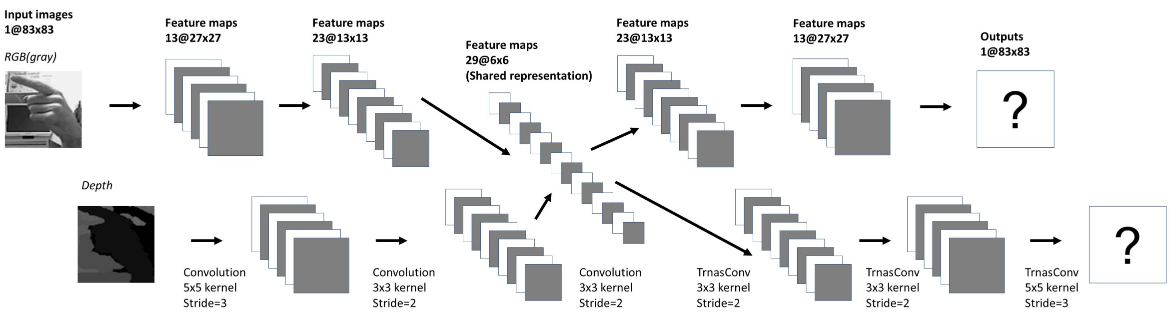

Figure 7: **The bimodal convolutional auto-encoder model that is used to learn shared multimodal representation.**
We simply take the CAE architecture that is introduced earlier (Figure 5) for each modaliy but force them to have a shared middle layer by adding the corresponding activation values. We then try to reconstruct the two images separately through two disjoint paths.

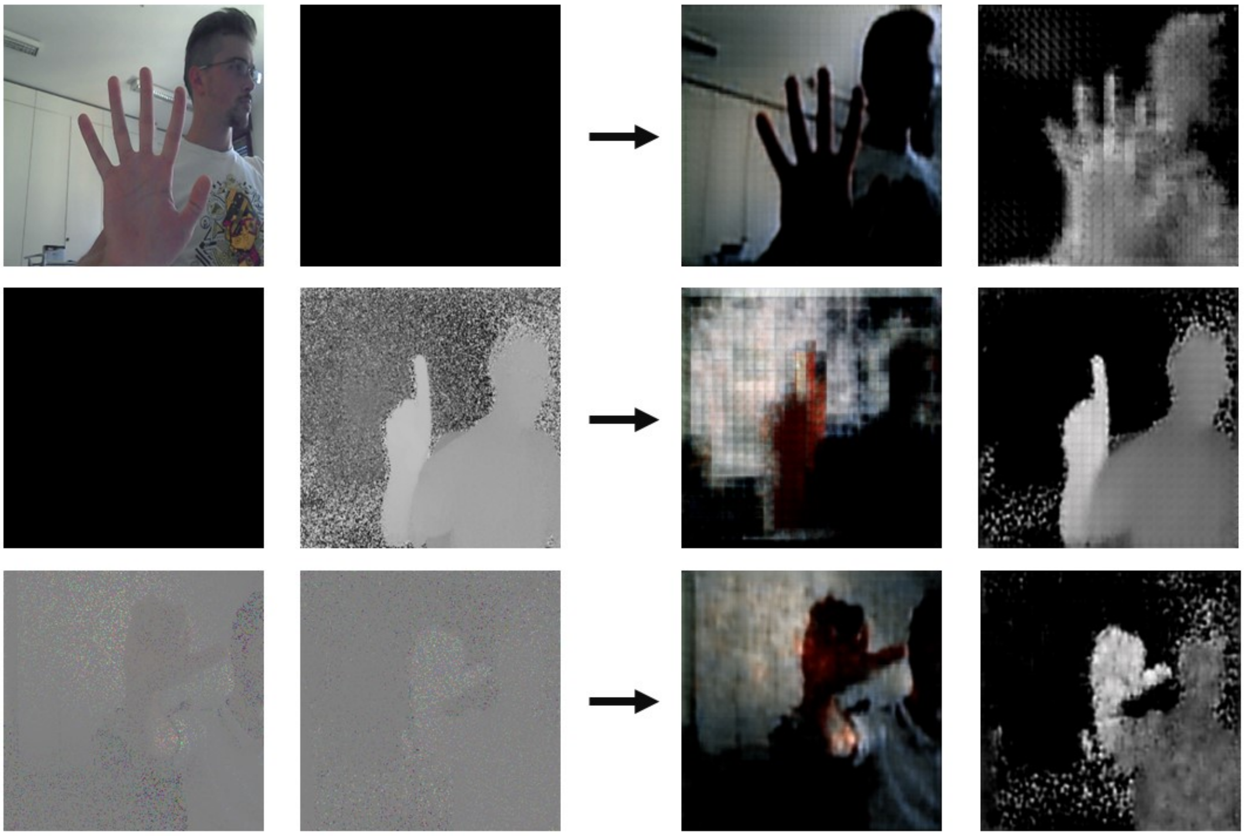

Figure 8: **Restore color and depth images from incomplete input information.**
Top) Only the color image is given.
Middle) Only the depth image is given.
Botttom) Both modalities are given but with little information (10% of pixels).

7.2 Transfer learning

References

[1] 17. M. Asadi-Aghbolaghi, A. Clapés, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, and S. Escalera. A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences. In *Automatic Face & Gesture Recognition (FG 2017) 2017 12th IEEE International Conference*, 2017.

[2] T. Baltrusaitis, C. Ahuja and L-P. Morency. Multimodal Machine Learning: A Survey and Taxonomy. In *CoRR*, vol. abs/1705.09406, 2017.

[3] C. Bregler and Y. Konig. "Eigenlips" for robust speech recognition. In *ICASSP*, 1994.

[4] L. Deng, G. Hinto and B. Kinsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In *ICASSP*, 2013.

[5] A. Droniou, S. Ivaldi, and O. Sigaud. A deep unsupervised network for multimodal perception, representation and classification. In *Robotics and Autonomous System*, 2014.

[6] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[7] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. In *IEEE Signal Processing Mag* Vol. 29, 2012.

[8] S. Ioffe and C. Szegedy. Batch normalization, accelerating deep network training by reducing internal covaraiate shift. In *CoRR*, vol. abs/1502.03167, 2015.

[9] A. K. Katsaggelos, S. Bahaadini and R. Molina. Audiovisual fusion: Challenges and new approaches. In *Proceedings of the IEEE*, vol. 103, 2015.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012.

[11] J. Masci, U. Meier, D. C. san, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *Artificial Neural Networks and Machine Learning–ICANN*, 2011.

[12] I. Matthews, T. F. Cootes, J. A. Bangham and S. Cox. Extraction of visual features for lipreading. In *PAMI*, 2002.

[13] A. Memo, L. Minto and P. Zanuttigh. Exploiting Silhouette Descriptors and Synthetic Data for Hand Gesture Recognition. In *STAG: Smart Tools & Apps for Graphics*, 2015.

[14] A. Memo and P. Zanuttigh. Head-mounted gesture controlled interface for human-computer interaction. In *Multimedia Tools and Applications*, 2017.

[15] S. Mitra and T. Acharya. Gesture recognition: A survey. In *IEEE Systems, Man, and Cybernetics*, 37:311–324, 2007.

[16] P. Molchanov, S. Gupta, K. Kim, and J. Kautz. Hand gesture recognition with 3D convolutional neural networks. In *CVPRW*, 2015.

[17] S. Moon, S. Kim and H. Wang. Multimodal transfer deep learning with applications in audio-visual recognition. In *NIPS MMML Workshop*, 2015.

[18] J. Nagi, F. Ducatelle, G. Di Caro, D. Ciresan, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L. Gambardella. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *Proceedings of the IEEE International Conference on Signal and Image Processing Applications*, 2011.

[19] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. Multiscale deep learning for gesture detection and localization. In *ECCVW*, 2014.

[20] J. Ngiam, A. Khosla, M. Kim, J. Nam, and A. Y. Ng. Multimodal deep learning. In: In *International Conference on Machine Learning*. 28. Bellevue, Washington, USA, 2011.

[21] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno and T. Ogata. Audio-visual speech recognition using deep learning. In *Applied Intelligence, vol. 42, no. 4, pp. 722–737, 2014.*

[22] E. Petahan, B. Bischoff, D. Bodoff, and N. M. Brooke. An Improved Automatic Lipreading System to enhance Speech Recognition. In *ACM SIGCHI*, 1988.

[23] L. Pigou, S. Dieleman, P. J. Kindermans, and B. Schrauwen. Sign language recognition using convolutional neural networks. In *Workshop at the European Conference on Computer Vision*, pages 572–578, 2014.

[24] G. Potamianos, C. Neti, J. Luettin and I. Matthews. Audio-visual automatic speech recognition: An overview. In *Issues in Visual and Audio-Visual Speech Processing, MIT Press*, 2004.

[25] N. Pugeault, and R. Bowden. Spelling It Out: Real-Time ASL Fingerspelling Recognition. In *Proceedings of the 1st IEEE Workshop on Consumer Depth Cameras for Computer Vision*, 2011.

[26] R. Socher, M. Ganjoo, H. Sridhar, O.Bastani, C. D. Manning, and A. Y. Ng. Zero-shot learning through cross-modal transfer. In *International Conference on Learning Representations (ICLR)*, Scottsdale, Arizona, USA, 2013.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. In *Journal of Machine Learning Research*, vol. 15, 2014.

[28] T. Starner, A. Pentland, and J. Weaver. Real-time american sign language recognition using desk and wearable computer based video. In *PAMI*, 20(12):1371–1375, 1998.

[29] J. Sung, I. Lenz, and A. Saxena. Deep multimodal embedding: Manipulating novel objects with point-clouds, language and trajectories. In *CoRR*, vol. abs/1509.07831, 2015.

[30] C. Szegedy, S. Ioffe, V, Vanhoucke. Inception-v4, inception-ResNet and the impact of residual connections on learning. In *CoRR*, vol. abs/1602.07261, 2016.

[31] V. Turchenko, E. Chalmers and A. Luczak. A deep convolutional auto-encoder with pooling - unpooling layers in Caffe. In *CoRR*, vol. abs/1701.04949, 2017.

[32] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. In *Machine Learning*, 81(1):21–35, 2010.

[33] Di Wu, Lionel Pigou, Pieter-Jan Kindermans, Nam Do-Hoang Le, Ling Shao, Joni Dambre, and Jean-Marc Odobez. Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition. In *Pattern Analysis and Machine Intelligence IEEE Transactions*, vol. 38, 2016.

[34] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski. Integration of acoustic and visual speech signals using neuralnetworks. in *IEEE Comm. Magazine*, pp. 65–71, 1989.