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

Hsieh Yu-Guan

August 6, 2017

Abstract

1 Introduction

2 Related Work

3 Presentation of Basic Network Architectures

4 Datasets and Preprocessing

4.1 Creative Senz3D

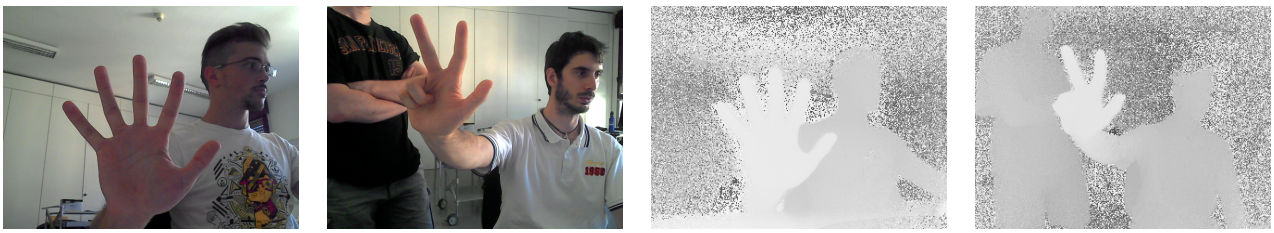

Figure 1: **Example images in the Creative Senz3D dataset.**
Left Two) Color images.
Right Two) Corresponding depth images.
All of the images are of size 480×640 and contain the the entire upper body of the subject.

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

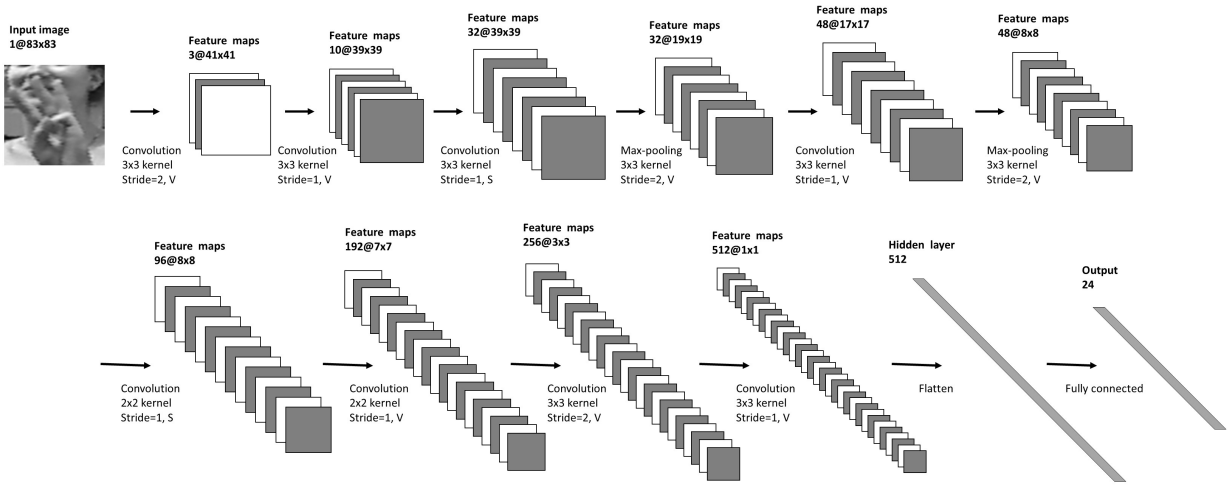

Figure 4: **CNN architecture used for the Finger Spelling dataset.**
The input of the nework is a one-channel image of size 83×83. It contains ten hidden layers. S stands for 'SAME' padding and V stands for 'VALID' padding (see text).

6.2 Convolutional auto-encoder

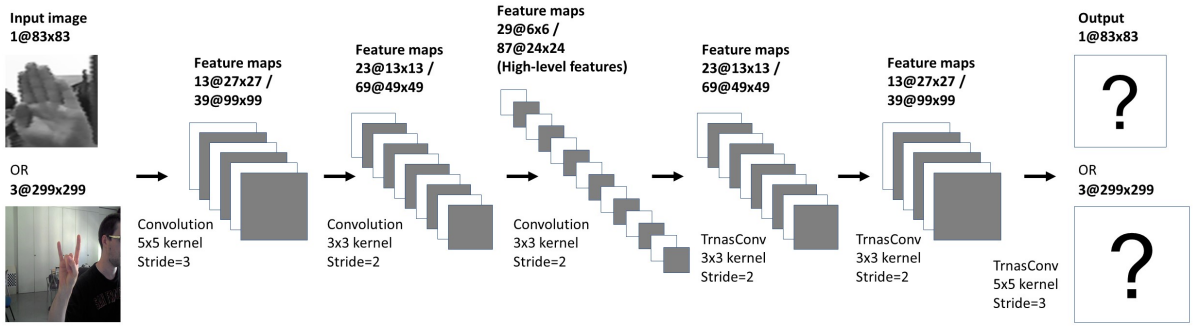

Figure 5: **Convolutional auto-encoder architecture with three convolutional layers and three tranposed convolutional layer.**
Activation values of the middle layer are taken as high-level features of the input image. Inputs of the network can be of different sizes. We only use valid paddings here.

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