audio-visual perception for recognition of human activities and object affordances for assistant robots.

May 29, 2016

1 network

In order to perform the detection of human activities, we have planned to use three networks that will detect human sub-activities based on different inputs. The Three inputs we consider are :

- The visual context at the proximity of both hands.
- The global audio recording.
- The human skeleton position and speed informations.

These inputs are going to be processed differently. The first two inputs (visual and audio) are processed by Convolution Neural Network (CNN) and the third input, is processed thanks to hand-engineered features. These processing, results in three types of descriptors, possibly describing sub activities.

Visual CNN

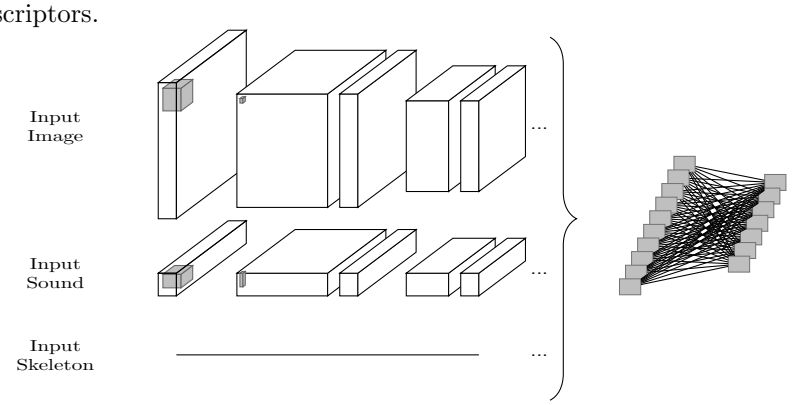
The first processing channel we describe is the visual one. It's input is composed by a crop of the scene centered and normalized on the hand it's also a three dimensional array which contains an RGB-image. This image is processed by a CNN that has been trained on ImageNet2. At the end of the network, we aim at having descriptors indicating which sub-activities are being performed, such as having a phone on the hand.

Audio CNN

The second processing channel is based on the audio recording. We extract a sequence of ?? seconds and classify it thanks to a CNN. Here also, the network is trained on a third party dataset and is intended to retrieve relevant descriptors also helping at identifying sub-activities. We think here at recognizing people speaking or sounds being produced by object manipulation.

Skeleton pose

The third and last processing channel we use, before aggregating all the aforementioned channels, is the skeleton processing. For this channel, we are considering using a *pose descriptor* as described in [2] resulting in spacio-temporal descriptors.



2 Dataset

In this section we present the different datasets used to train the four different parts of our network. Hence, we present the dataset used for image processing followed by the dataset used for sound processing, then the one for skeleton processing and, finally, the dataset that'll be used for training the RNN. The intuition beneath such needs is presented in section 1.

Image dataset

ImageNet[1] is a reference dataset in image processing. Many famous CNN architecture are based on this dataset such as AlexNet, GoogLeNet or NiNet. A contest is based on this dataset and the 2015'th edition of this contest has more than 450000 images of mean resolution of $482 * 415$ pixels organized in 200 labeled categories (such as airplane, ant, antelope, apple and axe) for object detection in images [1]. Though ImageNet 2015's contest doesn't include neither phones nor book, we can append them to our dataset thanks to the ImageNet main dataset consisting of over $14 * 10^6$ images [2]. We use this dataset to train our 2D image convolution neural network seen on section1.

For later uses, we might consider the video dataset consisting of 401 categories for scene classification also present on the ImageNet 2015's contest.

Audio dataset

skeleton dataset

[1] `http://www.image-net.org/challenges/LSVRC/2014/`
[2] `http://www.image-net.org/synset?wnid=n02992529#`

3 Testing

For testing purposes, we've build three datasets. We believe that the evaluation of human action recognition is increasingly harder on each of these datasets.

3.1 First set

The first set we recorded captured 5 actions : sitting down, reading, calling, drinking and sitting up. Each of these actions were captured based on the following scenario:

There is a camera behind a blank table seeing both the table and a chair behind it. A man comes in and sit on the chair. He'll then perform a set of three (non-sitting) actions. First, he'll read a book, secondly, he'll answer a phone call and thirdly he'll drink the content of a mug. The setup is such as the table is clear at the beginning of these action. When an action begins, the man puts the object he'll use on the table, then he uses it (reading, calling or drinking) and, at the end of the action, he puts back the object on the table before clearing it off. Once these three actions are performed, the man leaves the room. To see a video extract of this dataset you can watch a Youtube video[3].

As a technical matter, we've captured a 340p RGBD video stream at 15 fps and the audio stream at 28kHz. The depth image we use is the Xtion's one.

References

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[2] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2752–2759, 2013.

[3] `https://www.youtube.com/watch?v=VvPMcFAKO3U`