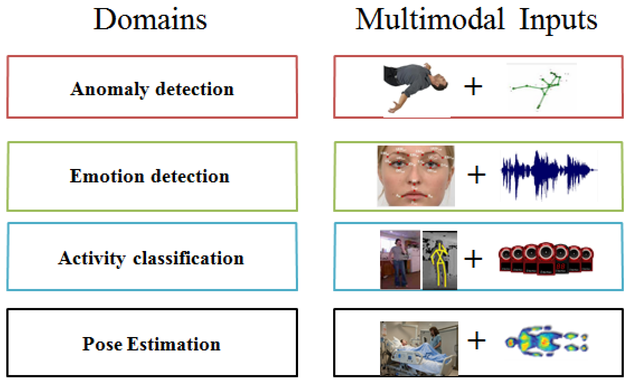
Multimodal Deep Emotional Activity Poses (M-DEAP)

Yash Chitalia[1] , Nayef Ahmar[2,3]

[1] Mechanical Engineering [2] Applied Physiology [3] Electrical and Computer Engineering

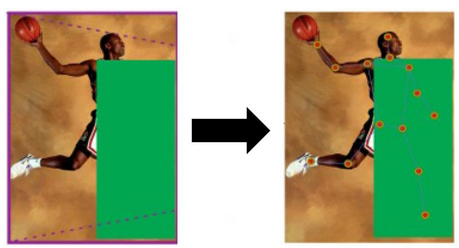
Github code

Motivation



- There is not enough manpower to serve baby boomers. Solution is robot health care assistant.
- Multiple problems need to be solved first, such as pose estimation under heavy occlusion, activity classification, emotion detection, and so forth.
- All these problems are of multimodal nature. This leads to a unifying theme, a domain free multimodal architecture.

Introduction

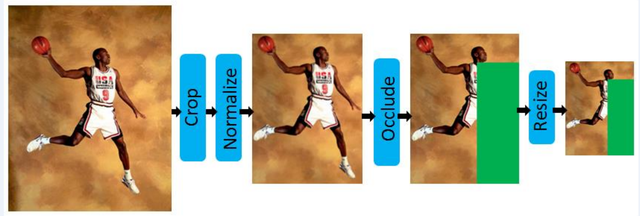


Goal : Apply Multimodal Deep Belief Network(DBN) and/or Multimodal Deep Boltzmann Machine(DBM) on RGB and depth images to estimate poses of occluded images.
– Learn joint "modality-free" representation.
– Infer missing modalities given some observed ones.
Method : Build a joint density model using a DBN/DBM
– Use states of top level hidden units as joint representation.
Dataset : we use the CAD-60 developed at Cornell[3] as our multimodal input source.
– Set of 60 videos and poses information of 5 subjects using Kinect sensor recording both RGB and Depth.

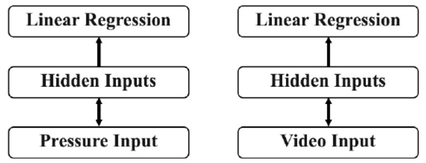
Method

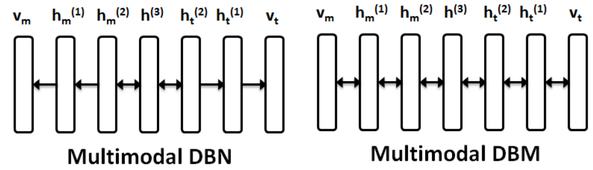
Preprocessing:

- Raw image is cropped by a bounding box.
- The joint positions are normalized w.r.t the width and height of the cropped image.
- Artificial occlusion is added to both RGB and depth images in a pseudo-random manner.
- Cropped image is resized to a fixed size (90 x 60).

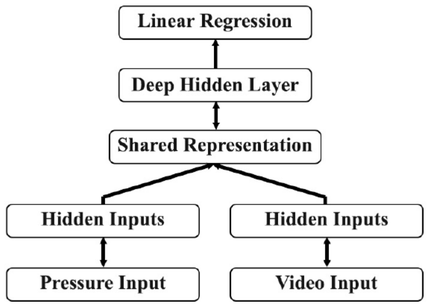


Architectures





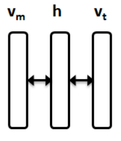
DBN: 4 hidden layers +1 linear regression.

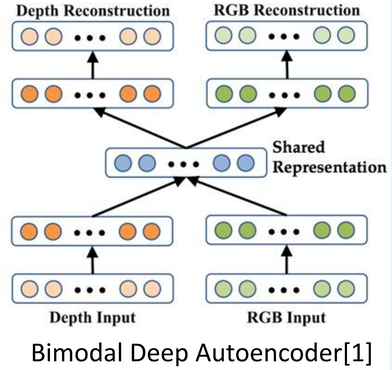


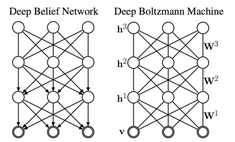
Multimodal DBN: 1 hidden layer per modality + hidden layer trained on shared representation.

Gaussian RBM

$$E(v,h) = \sum_i \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i,j} \frac{v_i}{\sigma_i} W_{ij} h_j - a^\top h$$







Bimodal Deep Autoencoder[1]

Preliminary Experimental Results

1-layer DBN(last) Hidden Units	Error	RGB	Depth
8100	Train	0.004	0.145
	Test	4e10	0.150
5400	Train	3e-4	0.073
	Test	8e12	1.024
2700	Train	0.015	0.002
	Test	3.7e11	200.919
1000	Train	0.043	0.016
	Test	623.785	0.755
30	Train	0.164	0.165
	Test	0.166	0.170

# epochs	DBN (secs)	DBM (secs)
1	276	499
3	793	1121
10	2615	3190

Performance speed DBN vs. DBM

Single (left) and 4 layer (right) DBN final layer with different number of hidden units

4-layers DBN (last) Hidden Units	RGB error	Depth error
750	0.061	0.018
500	0.057	0.016
100	0.066	0.041
30	0.106	0.063

Shallow M-DBN output	Layer-1/ Layer-2	1000/ 1000		
1000	Train	0.014		
	Test	0.208		
750	Train	0.025		
	Test	0.042		
500	Train	0.020		
	Test	0.071		
250	Train	0.032		
	Test	0.042		
30	Train	0.062		
	Test	0.065		

Future Work

- Test on deep – pose
- Run a larger dataset as opposed to a small sample
- Explore more broadly Multimodal DBM since it perform better in case of missing information.
- Work on other modalities

Acknowledgements

We thank Dr. Zsolt Kira for discussion, support, and guidance.

References

[1] Ngiam, Jiquan, et al. "Multimodal deep learning." Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011.
[2] Srivastava, Nitish, and Ruslan R. Salakhutdinov. "Multimodal learning with deep boltzmann machines." *Advances in neural information processing systems*. 2012.
[3] Sung, Jaeyong, et al. "Human Activity Detection from RGBD Images." plan, activity, and intent recognition 64 (2011).