Fully Convolutional Network for Head Detection with Depth Images

Diego Ballotta, Guido Borghi, Roberto Vezzani, Rita Cucchiara
University of Modena and Reggio Emilia, Italy

AImage Lab

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

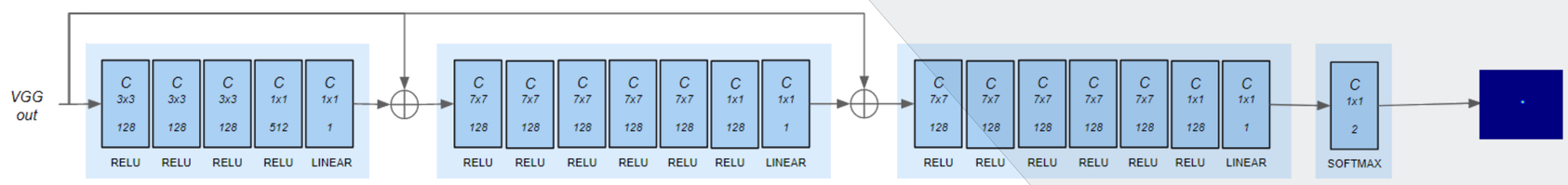
Abstract

In this paper, we propose a novel method for **Head Detection** on **depth images**, based on a deep learning approach. In particular, the presented system overcomes the classic *sliding-window* approach, that is often the main computational bottleneck of many object detectors, through a **Fully Convolutional Network**.

Even though in last decades many efforts have been conducted for head detection and localization with conventional light-visible cameras, **only few works tackle the problem of head detection on different types of images**, like *depth maps*.

Two public datasets, namely **Pandora** and **Watch-n-Patch**, are exploited to train and test the proposed network. Experimental results confirm the effectiveness of the method, that is able to **exceed all the state-of-art works based on depth images** and to run with **real time** performance.

Proposed Method



Input: depth maps (512x424 16-bit images, from *Microsoft Kinect One*)
Output: 64x53 probability map (as *bi-variate Gaussian* function)
Network Architecture: Fully Convolutional Network (inspired by [6])
Network Details: (ReLU + linear activation for each block) + softmax
Loss function: categorical cross-entropy

Final Head crop:

$$w, h = \frac{f_{x,y} \cdot R}{D}$$

Where $f_{x,y}$ are the horizontal and vertical focal lengths, R is the average value representing the width of a face (200mm) and D is the distance between the acquisition device and the head in the scene.

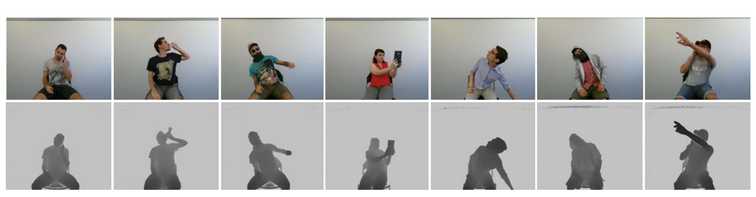
Datasets

Generally, head detection task with depth images **lacks datasets specifically created for the task**, containing a number of annotated sample that allow deep learning based approaches.

Pandora

This dataset has been presented in [1] and it has the following features:

- **250k frames** from **110 sequences**
- **22 subjects** (10 males and 12 females).
- Both **depth** and **RGB** frames and **skeleton** annotations frame by frame.
- Original tasks: **head and shoulder pose estimation**. Subjects can vary their head appearance wearing **garments** and objects like smartphones, tablets and plastic bottles that can generate head and body occlusions.



Watch-n-Patch

This dataset [2] is created for the modeling of human activities, comprising multiple actions in a completely unsupervised setting. It is composed of:

- **458 videos** with a total length of **230 minutes**
- **7 subjects** performing daily activities

Even if this dataset has not been explicitly created for head detection tasks, **it is a useful dataset to test head detection systems on depth images**, thanks to its variety in head poses, daily actions and subjects.

Experimental Results

For a fair comparison with literature methods that we consider as competitors (**works based on depth only** [3, 4, 5]), we split the *Watch-n-Patch* dataset as in the original work.

Moreover, the evaluation metric is the number of true positives, or rather the number of head detected. **A head is correctly detected only if**:

$$IoU(A, B) > 0.5$$

$$IoU(A, B) = \frac{Overlap\ Area}{Union\ Area} = \frac{|A \cap B|}{|A \cup B| - |A \cap B|}$$

Our method largely overcomes all other competitors, both in terms of true positive and false positive rates

Methods	Year	Method	TP	FP
Nghiem *et al.* [3]	2012	SVM	0.519	0.076
Chen *et al.* [4]	2016	LDA	0.709	0.108
Ballotta *et al.* [5]	2017	CNN	0.883	0.077
Our	**2018**	**CNN**	**0.964**	**0.036**

As far as it is concerned the frames per second, Ballotta *et al.* [5] method doubles our result, however the accuracy falls drastically.

Methods	True Positives	IoU	fps
Ballotta *et al.* [5]	0.956	0.806	0.238
Ballotta *et al.* [5]	0.717	0.552	31.5
Our	**0.984**	**0.789**	**16.79**

All tests have been carried on a **Intel i7-4790** CPU (3.60 GHz) and with a **NVIDIA GTX 1080 Ti**.

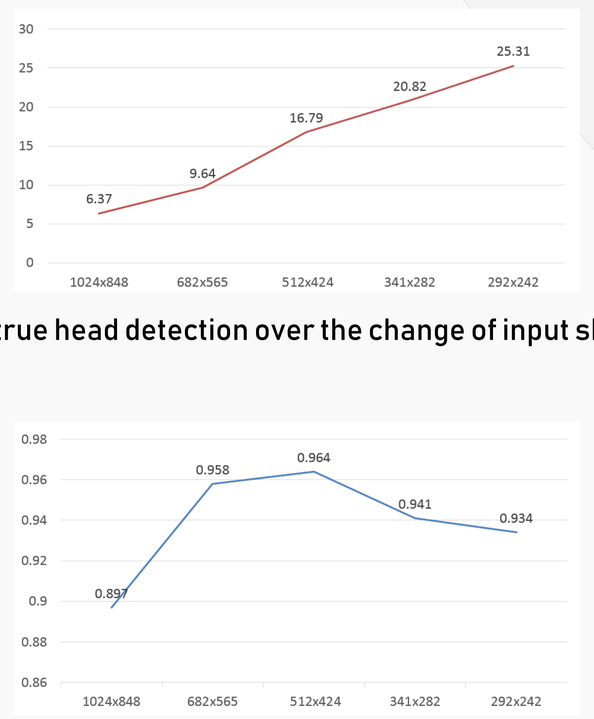
Deep models have been implemented and tested with **Keras** and **Theano** back-end.

Conclusion

In this paper, a novel method to directly detect and localize a head in depth images has been presented.

The proposed solution is based on a *Fully Convolutional Network* that is able to **output a probability map of the head locations**, given **only a depth map as input**.

Experimental results show the **accuracy**, the **reliability** and the **speed performance** of the framework on two public datasets.



Rate of true head detection over the change of input shape

Speed performance over the change of input shape (*fps*)

Download Pandora dataset

[1] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "*Poseidon: Face-from-depth for driver pose estimation*" (CVPR 2017)
[2] C. Wu, J. Zhang, S. Savarese, and A. Saxena, "*Watch-n-patch: Unsupervised understanding of actions and relations*" (CVPR 2015)
[3] A. T. Nghiem, E. Auvinet, and J. Meunier, "*Head detection using kinect camera and its application to fall detection*" (ISSPA 2012)
[4] S. Chen, F. Bremond, H. Nguyen, and H. Thomas, "*Exploring depth information for head detection with depth images*" (AVSS 2016)
[5] D. Ballotta, G. Borghi, R. Vezzani, R. Cucchiara, "*Head Detection with Depth Images in the Wild*" (VISAPP 2017)
[6] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "*Convolutional Pose Machines*" (CVPR 2016)

IAPR ICPR 2018
INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
In Beijing, China — August 20th-24th 2018