Fully Convolutional Network for Head Detection with Depth Images

Diego Ballotta, Guido Borghi, Roberto Vezzani and Rita Cucchiara
DIEF - University of Modena and Reggio Emilia
Via P.Vivarelli 10, 41125 Modena, Italy
Email: {name.surname}@unimore.it

***Abstract*—Head detection and localization are one of the most investigated and demanding tasks of the Computer Vision community. These are also a key element for many disciplines, like Human Computer Interaction, Human Behavior Understanding, Face Analysis and Video Surveillance. In last decades, many efforts have been conducted to develop accurate and reliable head or face detectors on standard RGB images, but only few solutions concern other types of images, such as depth maps. 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. 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.**

I. Introduction

In last decades, many efforts have been conducted for head detection and localization, one of the traditional computer vision research field. Specifically, many accurate and competitive solutions, used both in real world applications and in academic research, have been achieved on images taken by conventional light-visible cameras.

Recently, literature works have focused their attention on head detection in *wild* contexts, characterized by significant variations in illuminations, face expressions, pose and scales [1]. Moreover, the head, or the face, can be far from the acquisition device or turned away. A reliable head localization and detection is a key element and the first step for tasks based on faces, like head pose estimation, face recognition, human tracking, pedestrian detection and so on.

Only few works tackle the problem of head detection on different types of images, like *depth maps*, also know as range or 2.5D images. The latest wide spread of low cost, high quality and ready-to-use depth devices has encouraged the research community to investigate the use of this type of images [2], for various computer vision tasks, also including head detection [3].

Generally, the use of depth maps generates two benefits. The first is the reliability against light changes, a fundamental and needed feature in applications that are required to work also in absence or in particular conditions of light. An example is a driver's attention surveillance system, based on in-car images, that can be affected by severe light changes, due to different weather conditions or the night. The second benefit is that depth data could be exploited to deal with one of the main issue of many object detectors, caused by the unknown size of the target object, that obligates detector to run on different scale levels in respect to the original image [3]. Moreover, depth acquisition devices are now based on infrared light and not lasers, so they can be used in indoor human environments without particular constraints and limitations.

Fig. 1. Head detections on two sample frames taken from the *Watch-n-Patch* dataset [4]. This dataset contains daily activities performed in different environments and acquired by a depth sensor (*i.e. Microsoft Kinect One*). It is a challenging dataset for the head detection task, due to the presence of complex backgrounds, garments, extreme head poses and various occlusions.

Considering these elements, we propose a deep model to detect and localize heads in depth images. In particular, we investigate how to overcome the traditional *sliding-window* approach for the head detection task. The sliding-window technique is often the main computational bottleneck, due to the exhaustive search of heads on different levels of scale in the input image.

To test the presented system, two public datasets are exploited, namely *Watch-n-patch* [4] and *Pandora* [5] datasets, that contain both RGB and depth images of daily actions in various environments. Even if these datasets has not been created for head detection task, they are useful to test the proposed system due to the presence of complex backgrounds, different views and head poses and object interactions that produce head occlusions.

In this paper, we propose a head localization and detection method, working only on depth data. The presented system is based on a *Fully Convolutional Network* [6], designed to have good accuracy and real time performance. The first dataset exploited, *Watch-n-Patch*, is used both for the training and the testing phase of the network, while the second one, *Pandora*, is used only to test the presented method, performing then a cross dataset evaluation. Experimental results confirm the feasibility and the effectiveness of the proposed approach, also for real world applications.

This paper is organized as follows. Section II presents a summary of related literature works about head detection with RGB and depth images. Section III reports the details of the proposed method: in particular, the architecture of the network and the training and testing procedures are described. In Section IV experimental results and datasets are presented and finally conclusions and future work are drawn.

II. RELATED WORK

Head or face detection is a classic research topic for the Computer Vision community, and a great amount of detection systems have been developed for intensity images in last decades. The seminal work was proposed in [7], a boosting cascade framework for rapid object detection. Specific features have to be collected to increase the performance of the *Adaboost* classifier [8] in an uncontrolled environment, in which face can appear with very different poses. Furthermore, solutions based on different types of features, like HOG [9], SURF [10] and various classifiers like SVM [11] and Neural Networks [12] have been proposed. Also Deformable Parts Models (DPM) [13], [14] have been successfully applied for head and object detection on RGB images.

Recently, Convolutional Neural Networks (CNNs) have raised a great allure, due to their superior performance on various computer vision tasks, like object classification and regression and various generic object detectors based on CNNs have been proposed [15], [16]. In particular, *Faster R-CNN* [17] achieves a good compromise between computational efficiency and detection performance. Deep learning based methods often merge different tasks, like face detection, pose estimation and facial landmark localization [18].

All above reported methods use as input intensity images and only few works propose methods or approaches for head detection in depth images. Furthermore, the potentiality of the combination of depth maps and CNNs has not been completely investigated in the literature, yet. Recently, Chen *et al.* [19] proposed a circular head descriptor to classify each pixel in input images as belonging or not to a head. False positives are discarded and pixels are clustered through the analysis of depth data. In [20] 3D data are exploited to perform head detection for a fall detection framework. Only moving objects, obtained by a background subtraction, are detected and all possible head positions are searched on contour segments. Recently, a deep approach is exploited in [3], to obtain a classifier of head or non-head patches but this method relies on the classic sliding window approach and has not real time performance.

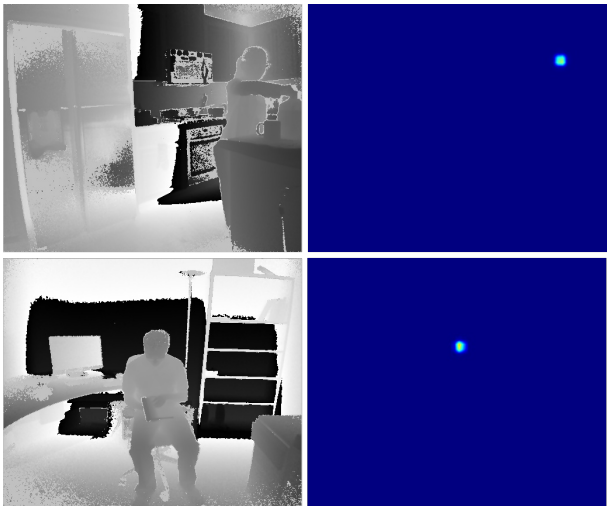


Fig. 2. Examples of ground-truth probability maps for head detection. For each input frame, a corresponding bivariate Gaussian distribution is created, centered on the head position.

Other works present in literature address the problem of only head localization: assuming the presence of only one subject into the scene, head coordinates are predicted. In [21] both head localization and orientation are predicted through a regression forests algorithm. In [5] a regressive CNN is exploited for head localization on depth images. In both cases, authors assume that the head is in the upper part of the body and at least partially centered in input frames. No temporal information are exploited to increase final accuracy.

Head detection task is often merged with the task of human or pedestrian detection. In these cases, specific features like EOH [22] or scale-invariant points (SIFT [23]) could be used. Human detection on RGB images is also based on specific local features, like *poselets* [24] and *edgelets* [25].

A human detector working on depth images is presented in [26], based on a 2D head contour model and a 3D head surface model. The final output is the entire body of users into the acquired scene. In [27] is proposed a multiple human detector. Input is represented by depth maps, and the method is based on a version of fast template matching algorithm. The output is the human body, obtained exploiting morphologic operators and segmentation techniques. Relational depth similarity features obtained from depth data are exploited in [28] in order to detect humans. Good accuracy and real time performance are achieved using *AdaBoost* classifier and integral images.

Finally, related to the diffusion of *Microsoft Kinect* device, in [29] is presented a method to detect the pose of one or more standing humans from single depth images, using the randomized decision trees algorithm.

III. HEAD DETECTION

The main goal of the proposed method is to detect the presence and estimate the positions of heads, given a single depth map. Good accuracy and real time performance are

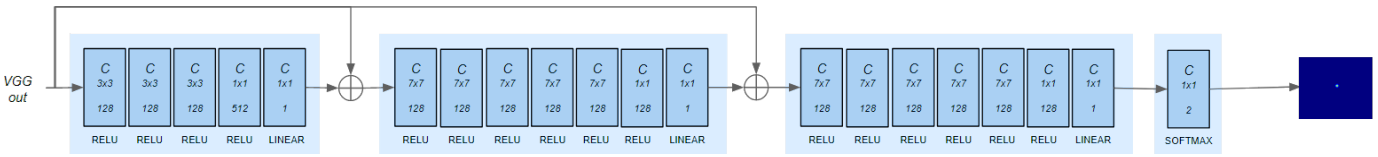


Fig. 3. Representation of the network architecture adopted in the proposed system. The first part consists of the four convolutional block of the VGG-19 network. Then, a cascade of different convolutional layers is added. Network takes as input 512×424 depth images.

also mandatory for real world applications, where promptness and reliability against light changes are essential. Taking inspiration from [30], we exploit a deep model, balancing speed and detection performance and avoiding the use of a sliding-window approach.

A. Depth Maps

We develop the proposed method starting from the analysis of the features offered by the depth acquisition device.
Both *Pandora* and *Watch-n-Patch* datasets have been acquired with the *Microsoft Kinect One*, a *Time-of-Flight* (ToF) device, that is able to recover the distance of each object inside the scene by measuring the time interval taken by an infrared light to reach the object itself and be reflected back to the camera. The adopted device can acquire both a depth map and the corresponding RGB image, with a spatial resolution of 512×424 and 1920×1080, respectively. It is also able to acquire data in real time (30 fps) and to capture depth information up to 8 meters, even if depth data are acquired only up to 5 meters. All depth data are reported in millimeters on 16 bits and are provided as a two dimensional array of pixels, like a gray-level image.
Due to the nature of ToF device, depth images acquired are generally affected by noise, especially along the edges, visible as random black spots.
As a result, in our system we use full resolution and 16 bit depth images as input data. We convert depth maps in traditional 8 bit gray-level images only for representation or visual inspection. Furthermore, we pre-process input images applying a 3×3 median filter to reduce the noise.

B. Network architecture

A *Fully Convolutional Network* is adopted and its architecture is depicted in Figure 3. We limit the depth of the final network to preserve and balance detection accuracy and speed performance. The network takes as input depth images of 512×424 pixels.
The first part of the network consists of the first 4 convolutional layers of the well-known VGG-19 network [31]. This part of the network provides the extraction of local visual features on the input data.
A cascade of convolutional layers follows the feature extraction part. The *ReLU* activation [32] is exploited in all hidden layers, with the exception of the last layer of each block, that use a *linear* activation. The final output layer is a convolutional layer that exploits a *softmax* activation over two filters, so that the *categorical cross-entropy* can be used as loss function for the network.

C. Training phase

The exploitation of pre-trained network weights is not possible since we use different input data. Thus, we train also the initial VGG-19 part from scratch. This training phase is done on a portion of *Watch-n-Patch* dataset. As first, input images are pre-processed with a rescale from 500 to 4500 mm, followed by the subtraction of the mean value, so their mean and variance are set to 0 and 1, respectively.
The ground truth for the network is artificially created as an image of shape 64×53 with a bi-variate Gaussian function, with a sigma empirically chosen ($\sigma = 25$ in our experiments), centered on the head located exploiting the dataset annotations. The network is trained to return as output a 64×53 probability map of head location over input, as depicted in Figure 2. Exploiting this approach, the proposed method is able to overcome the traditional sliding-window approach, producing benefits in speed performance and also detection accuracy, as reported in the following sections.
The network has been trained with a batch size of 10, a momentum value of 9^{-1} and a decay value of 5^{-4}. We exploit the *Stochastic Gradient Descent* (SGD) solver to solve the back-propagation problem, with an initial learning rate set to 10^{-2}. The use of other solvers that are able to automatically adjust the learning rate during the training, like *Adam* solver [33], produced worse results.
We apply data augmentation to the train data, in order to increase the number of training samples and avoid over-fitting phenomena, as reported in [34]: each input image is flipped, rotated and translated along the x and y axes.

D. Test phase

As above mentioned, the output of the deep model is a probability map. Therefore, a *Non-Maximum Suppression* is conducted over all output pixels with a threshold of 0.3, in order to isolate a single head location.
We obtain final bounding boxes of a certain width and height w, h of head detections as:

$$w, h = \frac{f_{x,y} \cdot R}{D} \tag{1}$$

where $f_{x,y}$ are the horizontal and the vertical focal lengths of the acquisition device, expressed in pixels (365.34 in our experiments), R is a constant value representing the average width of a face (200 mm in our experiments) and D is the distance between the acquisition device and a point in the scene and is computed as the mean value of a 15×15 pixel area, centered on detection coordinates. In this way, we are

TABLE I
RESULTS ON *Watch-n-Patch* DATASET FOR HEAD DETECTION TASK.

Methods	Year	Method	Features	True Positives	False Positives
Nghiem *et al.* [20]	2012	SVM	Modified HOG	0.519	0.076
Chen *et al.* [19]	2016	LDA	Depth-based Head Descriptor	0.709	0.108
Ballotta *et al.* [3]	2017	CNN	Deep	0.883	0.077
Our	**2018**	**CNN**	**Deep**	**0.964**	**0.036**

able to generate head bounding boxes accordingly with the distance of the subject from the camera, *i.e.* the dimension of the head.
Finally, thanks to the characteristic of the fully convolutional neural networks [6], the shape of input images is not a real issue, however use input shape too much different from the training shape may decrease network accuracy.

IV. EXPERIMENTAL RESULTS

Experimental results of the presented system are given using two public datasets, namely *Watch-n-Patch* and *Pandora* 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, even if in the last decades several datasets containing depth data were collected [35], [36], [21].
For a fair comparison with literature methods that we consider as competitors [19], [20], [3], we split *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:

$$IoU(A,B) > \sigma \tag{2}$$

$$IoU(A,B) = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|A \cap B|}{|A \cup B| - |A \cap B|} \tag{3}$$

where A, B are ground truth and predicted head bounding boxes, respectively. According to [19], the threshold σ is set to 0.5. Besides, we include also the computation time reported as frames per second (fps).

A. Pandora dataset

Pandora dataset is presented in [5] and it is composed of about 250k frames from 110 sequence acquired from 22 subjects (10 males and 12 females). It contains both depth and RGB frames and skeleton annotations frame by frame. This dataset is entirely acquired with the *Microsoft Kinect One* device and is created for the head and shoulder pose estimation tasks. It is a challenging dataset, subjects can vary their head appearance wearing prescription glasses, sun glasses, scarves, caps and objects like smartphones, tablets and plastic bottles can generate head and body occlusions. Moreover, subjects perform extreme poses with both head and shoulders.

TABLE II
RESULTS ON *Pandora* DATASET FOR HEAD DETECTION TASK. IN PARTICULAR, WE REPORT BOTH BEST SPEED PERFORMANCE AND BEST DETECTION RATE OF THE METHOD PRESENTED IN [3]. OUR METHOD OVERCOMES COMPETITOR BOTH IN ACCURACY THAN IN COMPUTATION TIME.

Methods	True Positives	IoU	fps
Ballotta *et al.* [3]	0.956	0.806	0.238
Ballotta *et al.* [3]	0.717	0.552	31.5
Our	**0.984**	**0.789**	**16.79**

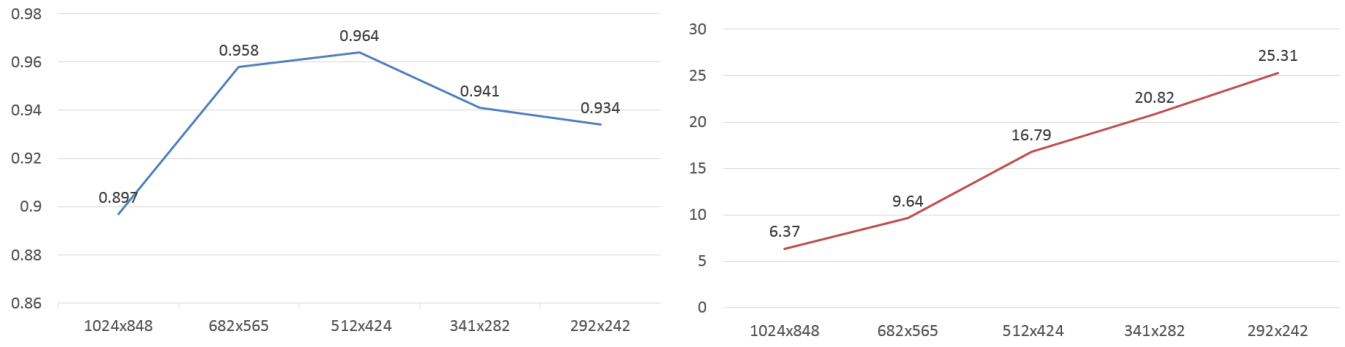
B. Watch-n-Patch

Wu *et al.* presented this dataset in [4]. It 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, split into 7 subjects performing daily activities in complex and various environments. Like *Pandora*, it is collected with *Microsoft Kinect One* sensor. Moreover, skeleton data are provided as ground-truth annotation. 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.

C. Quantitative evaluation

In this section, experimental results of the proposed system are reported. For a fair comparison, we report methods only based on depth data. As mentioned above, we compare our method with three recent works present in the literature [20], [19], [3]. For this first comparison, we exploit the *Watch-n-Patch* dataset, split in the same way of the original work [19], the test subset consists of around 2785 images. This subset has been chosen by authors due to the presence of scenes with a general good background quality, required by [20]. [1]
Table I shows a comparison against different competitors on the *Watch-n-Patch* dataset. Our method largely overcomes all other competitors, both in terms of true positive and false positive rates.
In Table II is shown a comparison on a subset of *Pandora* dataset, between our method and [3]. For the competitor, we implement the method and run the algorithm with two kernel size (3×3 and 45×45), to have best performance in detection accuracy and computation speed, as reported in [3]. Results achieved identify our method as best in term of true positive. As far as it is concerned the frames per second, [3] method doubles our result, however the accuracy falls drastically.
Experimental results with other head detection methods based on depth data [21], [5] are not feasible, since a specific context – *i.e.* a person facing the acquisition device – for acquired

[1]The test split is not reported in the original work. Authors have sent us the split by a private message.



(a) (b)

Fig. 4. The graph (a) show the rate of true head detections over the change of input shape (reported in x axis). Graph (b) shows the related speed performance expressed as frames per second.

scene is strictly required. As reported in Section III-D, a fully convolutional network can have inputs of different size. In Figure 4 the rate of true positives (left) and frames per second (right) are reported varying the dimension of input depth images.

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* [37] and *Theano* [38] back-end.

V. 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.

The flexibility of our approach allows a variety of future works. Multiple head detection will be investigated. Also the acquisition of a new dataset, specifically designed and collected for the head detection task is strongly required, due to the lack of data for this task. Finally, future extensions will be related with the improvement of speed performance also on low power GPUs and mobile platforms.

Acknowledgment

This work has been carried out within the project "FAR2015 Monitoring the car drivers attention with multisensory systems, computer vision and machine learning" funded by the University of Modena and Reggio Emilia. We also acknowledge the CINECA award under the ISCRA initiative.

References

[1] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: past, present and future," *Computer Vision and Image Understanding*, vol. 138, pp. 1–24, 2015.

[2] H. Sarbolandi, D. Lefloch, and A. Kolb, "Kinect range sensing: Structured-light versus time-of-flight kinect," *Computer Vision and Image Understanding*, vol. 139, pp. 1–20, 2015.

[3] D. Ballotta, G. Borghi, R. Vezzani, and R. Cucchiara, "Head detection with depth images in the wild," in *Proceedings of the International Conference on Computer Vision Theory and Applications*, 2018.

[4] C. Wu, J. Zhang, S. Savarese, and A. Saxena, "Watch-n-patch: Unsupervised understanding of actions and relations," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[5] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "Poseidon: Face-from-depth for driver pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[8] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005.

[10] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer vision–ECCV 2006*, pp. 404–417, 2006.

[11] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[12] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 1, pp. 23–38, 1998.

[13] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Real-time high performance deformable model for face detection in the wild," in *Biometrics (ICB), 2013 International Conference on*. IEEE, 2013, pp. 1–6.

[14] J. Yan, Z. Lei, L. Wen, and S. Z. Li, "The fastest deformable part model for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2497–2504.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[16] T.-H. Vu, A. Osokin, and I. Laptev, "Context-aware cnns for person head detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2893–2901.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[18] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[19] S. Chen, F. Bremond, H. Nguyen, and H. Thomas, "Exploring depth information for head detection with depth images," in *Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on*. IEEE, 2016, pp. 228–234.

[20] A. T. Nghiem, E. Auvinet, and J. Meunier, "Head detection using kinect camera and its application to fall detection," in *Information*

Fig. 5. Final output of the presented system. In the first and third rows are reported depth frames with the visualization of the predicted probability maps (blue is used for low values, colors tending to red for high probabilities). In the remaining rows, depth maps with head bounding boxes predictions (green) and ground truth annotations (blue) are reported.

Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on. IEEE, 2012, pp. 164–169.

[21] G. Fanelli, J. Gall, and L. Van Gool, "Real time head pose estimation with random regression forests," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 617–624.

[22] K. Levi and Y. Weiss, "Learning object detection from a small number of examples: the importance of good features," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. IEEE, 2004, pp. II–II.

[23] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[24] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1365–1372.

[25] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 90–97.

[26] L. Xia, C.-C. Chen, and J. K. Aggarwal, "Human detection using depth information by kinect," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*. IEEE, 2011, pp. 15–22.

[27] M. H. Khan, K. Shirahama, M. S. Farid, and M. Grzegorzek, "Multiple human detection in depth images," in *Multimedia Signal Processing (MMSP), 2016 IEEE 18th International Workshop on*. IEEE, 2016.

[28] S. Ikemura and H. Fujiyoshi, "Real-time human detection using relational depth similarity features," in *Asian Conference on Computer Vision*. Springer, 2010, pp. 25–38.

[29] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[30] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[32] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[33] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[35] A. D. Bagdanov, A. Del Bimbo, and I. Masi, "The florence 2d/3d hybrid face dataset," in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. ACM, 2011, pp. 79–80.

[36] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "3d constrained local model for rigid and non-rigid facial tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.

[37] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[38] R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, A. Belopolsky *et al.*, "Theano: A python framework for fast computation of mathematical expressions," *arXiv preprint*, 2016.