”Look At This One” Detection sharing between modality-independent classifiers for robotic discovery of people

Joris Guerry[1], Bertrand Le Saux[1] and David Filliat[2]

***Abstract*— With the advent of low-cost RGBD sensors, many solutions have been proposed for extraction and fusion of colour and depth information. In this paper, we propose new different fusion approaches of these multimodal sources for people detection. We are especially concerned by a scenario where a robot evolves in a changing environment. We extend the use of the Faster RCNN framework proposed by Girshick *et al.* [1] to this use case (i), we significantly improve performances on people detection on the InOutDoor RGBD People dataset [2] and the RGBD people dataset [3] (ii), we show these fusion handle efficiently sensor defect like complete lost of a modality (iii). Furthermore we propose a new dataset for people detection in difficult conditions: ONERA.ROOM (iv).**

I. INTRODUCTION

When exploring an unknown place, a robot can find itself in very different situations. Because of these uncertain conditions, different sources of information may be used to ensure greater detection robustness. In the context of computer vision, a colour camera (Red-Green-Blue) can be coupled to a so-called ”depth map” camera providing additional spatial information. Kinect or Xtion cameras are examples of RGBD sensors providing all of this information to the user. The low cost of these devices and their plug’n’play design have raised a recent interest for the scientific community. The nature of the information provided by each built-in sensor is different and can be complementary. For this reason, it is important to cleverly merge this information to improve classification performance and to be robust to specific failure conditions for each modality. In particular, for a robot equipped with such sensor (Figure 1) in an exploration scenario, situations such as a dark room can suppress the RGB modality, while sunny areas can suppress the depth modality.

In this paper, we focus on people detection because of its importance in many scenarios for domestic or search-and-rescue robotics. Many methods have been proposed for this detection task on RGB images, such as Histogram of Oriented Gradients for people detection [4] and Deformable Part Models [5] to name but a few. The use of a second modality like depth permits to be less dependant to colour and texture features and instead to focus on global geometric shape or object instance separation. Using such a multimodal RGBD approach, Gupta *et al.* [6] proposed a method based on the Girshick *et al.* Region-CNN [7].

[1]Joris Guerry and Bertrand Le Saux are with ONERA The French Aerospace Lab, F-91761 Palaiseau, France firstname.name@onera.fr

[2]David Filliat is with the ENSTA ParisTech, U2IS, Inria FLOWERS team, Université Paris-Saclay F-91762 Palaiseau, France david.filliat@ensta-paristech.fr

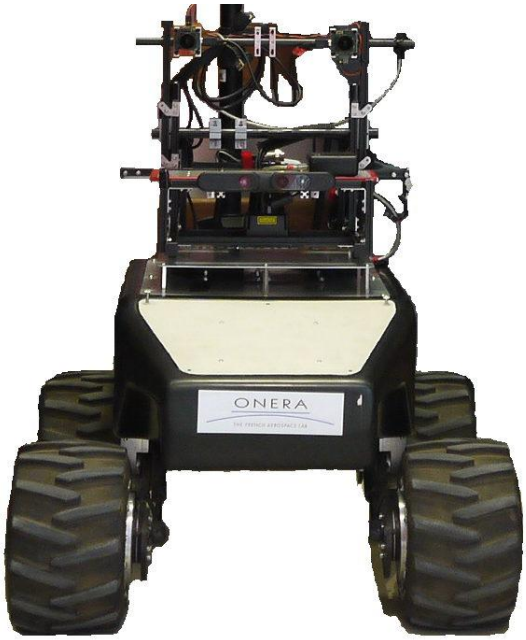

Fig. 1. Our mobile robot used for ONERA.ROOM acquisitions

They use an independent Region Of Interest (ROI) proposal module and then extract features of the region through a convolutional neural network (CNN). These features are subsequently classified by a support vector machine (SVM). The major contribution of their work is an efficient encoding of the depth information: HHA (Horizontal disparity, Angle of normal vector to gravity). However, the ROIs proposition phase as well as the HHA processing are time-consuming and the SVM makes the training process complex. Aware of this complexity problem of HHA encoding, Eitel *et al.* proposed in [8] to make a simple and fast rendering of the depth map, transforming the spatial distance into a colour information with the Jet colormap. The authors proposed to merge three CNNs, an RGB expert, a depth expert and an optical flow expert, replacing the last layer of each expert by a common fully-connected fusion layer and appending a final global softmax layer. The evolution of this work led to the use of a CNN module that weights the outputs of each expert on the fly, rather than learning the combination statically. This module, called Gating network [2], is based on features extracted at the last levels of each expert. Thus, with both class probability vector, a final vector is obtained as a weighted sum of the coefficients proposed by the Gating Network. Therefore, an expert can overwrite the information

of others if the Gating Network weights it sufficiently. Since the neural network used is non-linear, this kind of decision can lead to the complete loss of information of one of the experts.

Another multimodal fusion strategy was proposed by Hazirbas *et al.* with FuseNet [9], where the intermediate tensors of the depth expert neural network are concatenated in the RGB expert network. The convolution kernels of the depth expert remain independent of the colour, but the convolution kernels of the RGB expert must now process the depth information. If the resulting activation of the convolutional filters is not sufficient for at least a single modality, it is possible that the total activation is not sufficient for the macro network, locally blocking the information. For example, if the RGB image is too dark, only the depth expert should be able to express itself, which is not possible here. Badrinarayanan *et al.* [10] applied a new encoder-decoder CNN structure for semantic segmentation on a RGBD dataset showing better results with only RGB modality than former RGBD based methods. But this dataset was not presenting hard luminosity condition.

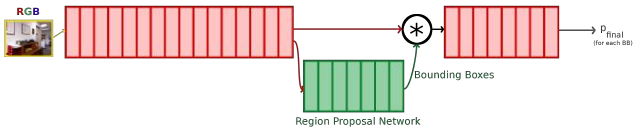

Fig. 2. Fr RCNN architecture [1]

However, some efficient methods have been also proposed for single modality. For range images [11] computes feature on the point cloud which is processed from the depth acquisitions. For RGB-only data, object detection is a very active domain. The R-CNN adapted by Gupta for RGBD has already been outperformed by Fast RCNN [12]. This is a deep network object classification method, also used in [2], which uses an ROIs proposal module independent of the classification network. In following works, the Region Proposal Network (RPN) has been added to a new version of the method: Faster RCNN [1] (abbreviated to Fr RCNN). The RPN provides regions of interest directly within the CNN architecture. The CNN ROIs provided by the RPN are then extracted from the 5^{th} convolution layer output to be classified by the second part of the network (see Figure 2). In this paper, we study how multimodal RGBD information can be exploited in the frame of the Fr CNN approach. Indeed, more than a particular neural network, Fr RCNN is a concept that can be applied to many structures of CNN. Yet, the Fr RCNN is just made for a single source of information and, to our knowledge, has not been structurally adapted to the use of multimodal sources. We will show that these approaches significantly improve performances on people detection on the InOutDoor RGBD People dataset [2] and the RGBD people dataset [3] and that these fusion handle efficiently sensor defect like complete lost of one of the modalities. We also propose a new dataset, ONERA.ROOM, containing more challenging detection conditions acquired in a mobile robotics exploration scenario.

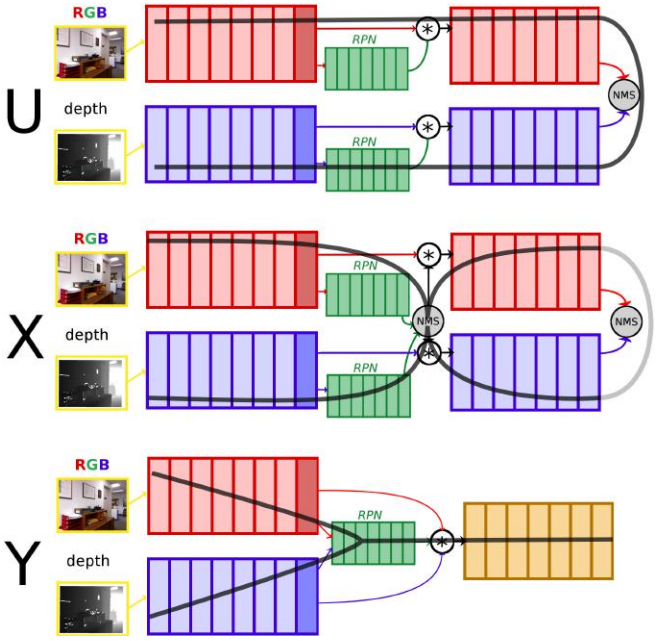

Fig. 3. Different fusion architectures, from top to bottom: U-fusion (NMS with the expert classifier outputs), X-fusion (NMS on the RPN outputs and NMS on the classifier outputs), Y-fusion (working on the concatenation of the tensors of each semi-expert).

II. JOINT-MODE APPROACH FOR OBJECT DETECTION

We propose multiple approaches for fusion of information from both modes: RGB and depth. First, we build single mode experts based on the effective region proposal module of Fr RCNN [1]. We show that this architecture can also be successfully applied to depth data. Second, the objective of fusion is to be able to adapt to realistic data in which the conditions of luminosity can completely degrade the performances of the RGB expert (in dark environment, blurry, smoky, ...) or of the depth expert (outdoor, long distance, ...). This is why the approaches we propose try to keep the experts independent (i.e avoid hybrid architectures) while allowing them to help each other.

In this paper, NMS is referring to Non-Maximum Suppression, recently called GreedyNMS in [13] as opposed to learnt NMS. This post-processing sorts and selects the best object detections among proposals in order to only keep one detection per object.

From now on, our different architectures are named RGBD RCNN (see Fig. 3) with the following variants.

- The **U** architecture is a naive version with two parallel networks. Thus, the two streams of information are only merged at the end of the detection. The fusion step is performed by NMS. This simple approach allows to have a better recall than single experts since one network can make up for the failing one but can lead to worse precision. This strategy can be applied to every computer vision technique that proposes bounding boxes.

- The **X** architecture is more subtle. Placing a common NMS after the RPN makes it possible to share the ROIs before the classification process of each expert. This pooling of detections allows an expert to share its detections with the other expert : "Look at this one !". Thus, class-blind object detections from both RGB and depth can be classified by the two experts meanwhile the redundant proposed regions are handled by the final NMS.
- The **Y** architecture aims to use only one RPN, taking as input the 5^{th} convolution outputs of both experts concatenated in a single tensor. The underlying assumption is that a RPN which gets both RGB and depth inputs will be able to predict better ROIs. However, the RPN feature space is now twice as big as before and thus training is more complex. So, longer time is required for optimisation. A second benefit of such a fusion is that it results in a lighter architecture with a single classifier, so less parmeters to optimize and faster prediction times.

The U and X architectures are structurally independent: RGB and depth experts are trained separately and it is only at test time that ROIs are shared. They are thus incremental approaches: a new sensor modality could be added without retraining the existing ones. On the contrary, the Y architecture would require a new training with all sources. The advantage is that it would also be able to learn cross-modality features.

In this paper, our new method uses Fr RCNNs based on the VGG16 [14] network[1]. Each training is done with stochastic gradient descent for 10,000 iterations with a constant learning rate of 0.001 through Pytorch framework. Parameters are initialized with pre-trained weights on the ImageNet [16] dataset. At test time, each expert runs at 3 frames per second (fps) and the X-fusion runs at 5 fps.

III. RESULTS

Fig. 4. Examples of predictions with models trained and tested on RGBD People dataset [17], from left to right : RGB expert, depth expert, X-fusion. The depth expert was able to find all the people in the image but the X-fusion propose better bounding box alignments with the ground truth.

We compare our approach with state-of-the-art methods on two public RGBD datasets for people detection: RGBD People in part III-A and the more recent InOutDoor dataset with a moving camera in part III-B. We also run experiments of people detection with our robotic platform in even more challenging set-ups, which yields the new dataset we deliver: ONERA.ROOM (in part III-C). For performance evaluation we use standard metrics of object detection used in similar contexts [2]: Average Precision (AP), Equal Error Rate (EER), Intersection-over-Union (IoU) and the harmonic mean of the precision/recall pair (F1).

[1]Fr-RCNN was also implemented with Residual Networks [15], leading to improvement over VGG networks on many datasets but at a large computational cost

A. RGBD PEOPLE DATASET

We first test our method on the RGBD People [3] dataset. This set of over 3000 *RGBD* images was acquired with 3 static cameras in a large hall with stable lighting conditions. By reproducing the experiment in [2], we produced 5 random sets (70 % training / 30 % test) and give the averaged results in Table I, ignoring cases of occlusions . We consider a correct detection if $IoU > 0.6$. The method Fr RCNN allows a significant gain on the method proposed in [2] (+9.1 points) and the X-fusion slightly increases this result (+0.2 points). An illustration of these results is shown in Figure 4.

TABLE I
EER ON RGBD PEOPLE [17] DATASET FOR SEVERAL DETECTORS. HOD IS SHORT FOR HISTOGRAM OF ORIENTED DEPTHS AND HGE FOR HIERARCHICAL GAUSSIAN PROCESS MIXTURES OF EXPERTS.

Method	Source	EER
HOD [18]	*D*	56.3
HGE [18]	*RGBD*	87.4
Gating Net. [2]	*RGBD-Optical flow*	89.3
Fr RCNN [1]	*D*	98.3
Fr RCNN [1]	RGB	98.4
RGBD RCNN U	*RGBD*	98.4
RGBD RCNN Y	*RGBD*	98.3
RGBD RCNN X	*RGBD*	**98.6**

B. INOUTDOOR RGBD PEOPLE DATASET (IOD)

TABLE II
PERFORMANCE ON THE IOD DATASET (REFERENCE SETS [2]) FOR DIFFERENT DETECTORS.

Method	Source	Precision/Recall	AP	EER	IoU	F1
Gating Net. [2]	*RGBD*	–/81.1	80.4	–	–	–
Fr RCNN [1]	RGB	**73.2**/94.2	91.9	90.1	**80.3**	**82.4**
Fr RCNN [1]	*D*	46.9/85.9	84.0	84.8	79.4	60.7
U	*RGBD*	45.6/96.4	**94.4**	92.1	**80.3**	61.9
X	*RGBD*	43.5/**96.5**	94.3	**92.4**	79.9	60.0
Y	*RGBD*	59.4/93.3	90.2	90.1	80.2	72.6
U *	~~*RGB*~~*-D*	46.9/85.9	84.0	84.8	79.4	60.7
X *	~~*RGB*~~*-D*	45.9/85.9	84.1	84.8	79.4	59.9
Y *	~~*RGB*~~*-D*	n.a/0	n.a	n.a	0	n.a

The more challenging IOD [2] dataset is obtained by an RGBD camera embedded on a mobile robot which evolves among people in motion. It has sharp changes in brightness (indoor / outdoor) and consists in 4 sequences (8305 images) captured at different times of the day with variations in ambient light. We consider a detection is correct if $IoU > 0.6$. The Fr RCNN [1] alone, (RGB or D) already achieves a

better score than the previous state of the art [2], with a gain of 11.5 points on the AP. This important gain can be partially explained by a better recall (detection rate) thanks to the RPN. Fusion strategies further improve the score by 2.5 points, raising AP to 94.4%. A classification example on IOD is available on the left column of Figure 5.

We notice that X and U strategies have better performance than Y. The training is more delicate for the Y strategy. Indeed, the task is more complicated because the feature space is bigger for the single RPN/classifier. Moreover, this type of hybrid network can not handle the loss of one of the sources as shown by the total absence of detections for Y * in the experiments with ablation of the RGB modality (see Table II). On the contrary, U * and X * react well by equalizing the performances of the depth expert alone.

Fig. 5. Examples of predictions with models trained on IOD [2] and tested on IOD and ONERA.ROOM datasets. It is interesting to notice that the X-fusion (left column) has a new ROI. This is allowed by the intermediate NMS, post-RPN: the depth expert found an object and was not able to classify it, however, the RGB expert did not initially find this object but was finally able to classify it. The U-fusion (right column) does not allow this pre-classification detection exchange but allows to reorder, by score, the ROIs during the final NMS. In this example, the ROI from the depth expert is dismissed in favor of the RGB expert central ROI, which allows the depth expert to provide a previously ignored ROI (in yellow).

C. ONERA.ROOM

We now propose a new and more challenging dataset in a robotic exploration scenario.

Robot framework: The experimental set-up consists in a four-wheeled Robotnik Summit XL (cf. Fig. 1) equipped with a RGBD camera (Asus Xtion in its last version, but formerly Kinect v1 and Realsense camera). The sensors are linked to an embedded computer and wi-fi transmitter for remote data processing.

Dataset: ONERA.ROOM is a new data set with 27 sequences acquired by various RGBD sensors (Kinect v1, RealSense and mainly Xtion), embedded on a remote-controlled mobile robot. 23 sequences containing people have been labelled, representing 27201 ROIs of people distributed in 35379 images. Oriented to robotic scenarios for search and rescue, the ONERA.ROOM dataset includes sequences acquired in the dark, sequences blurred by the movement of the robot and cases of unconscious people on the ground. It has three main sets of increasing difficulty level: "Easy", "Average" and "Hard", and is made publicly available for research purposes on our website[2].

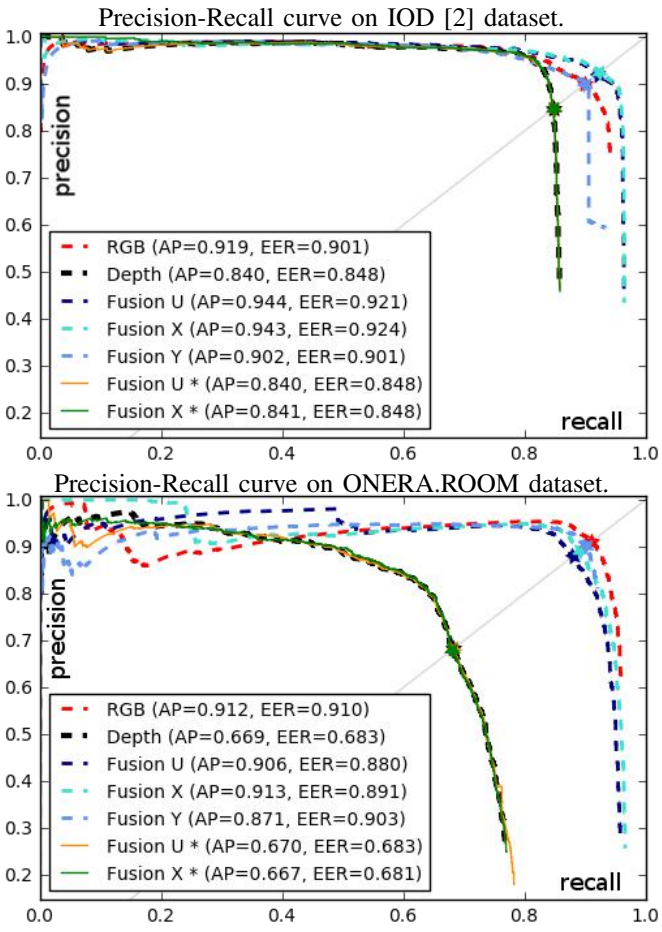

Fig. 6. The asterisk means that the source RGB was unavailable (black image). U and X-fusions behave well against this defect (see fusion curve U * and fusion X *) which fall back to the performance level of the depth expert. The X-fusion has a better AP than the RGB expert. Although the EER of the Y-fusion is good, this strategy is unable to make any detection in the event of loss of the RGB source.

Experiments: To impose a statistical independence between the training data and the test data, we used the models trained on the IOD *train* set and apply them to ONERA.ROOM. All the quantitative results on ONERA.ROOM refer to the *Easy* set. We consider a detection is correct if $IoU > 0.5$. The trends are similar to the experiments on IOD (see Figure 6 and Table III). The U and X fusions are better than the RGB expert and as good as the depth expert in the absence of light. A U-fusion classification sample is

[2]http://jorisguerry.fr/ONERA.ROOM

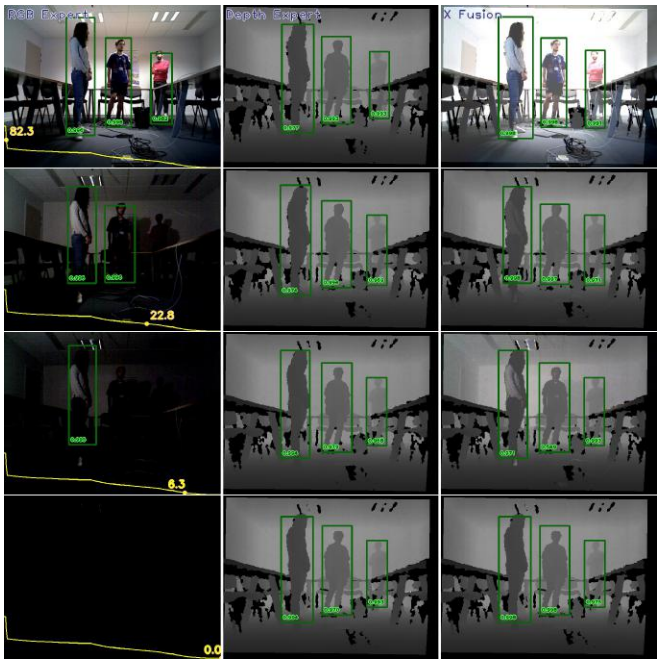

Fig. 7. Influence of decreasing luminosity. ONERA.ROOM propose a static sequence where the only changing factor is ambient light. The yellow curve indicate the mean pixel intensity. First column shows true positive detection of RGB expert, second column concerns depth expert and third column are the X-fusion detections. The last column images are made from the RGB and depth mean image for illustration purpose only.

TABLE III
PERFORMANCES ON THE *ONERA.ROOM* DATASET, "EASY" SET, FOR VARIOUS DETECTORS TRAINED ON IOD [2].

Method	Source	Precision/Recall	AP	EER	IoU	F1
Fr RCNN [1]	RGB	61.0/96.1	91.2	**91.0**	**72.8**	74.6
Fr RCNN [1]	depth	25.6/76.9	66.9	68.3	65.3	38.5
U	*RGBD*	26.7/95.8	90.6	88.0	71.1	41.8
X	*RGBD*	25.7/**96.6**	**91.3**	89.1	71.7	40.7
Y	*RGBD*	**81.1**/92.1	87.1	90.3	71.7	**86.3**
U *	~~*RGB*~~-*D*	18.2/78.3	67.0	68.3	65.3	29.5
X *	~~*RGB*~~-*D*	25.0/77.0	66.7	68.1	65.3	37.8
Y *	~~*RGB*~~-*D*	n.a/0	n.a	n.a	0	n.a

available in the right column of Figure 5. In the case of a rescue mission such an approach will be more robust to unpredictable, degraded conditions. A video illustrating several conditions is available on the website of the ONERA.ROOM dataset. Figure 7 illustrates RGB expert, depth expert and X-fusion behaviours facing luminosity reduction: when the environment is too dark for the standard RGB expert, the depth one is able to compensate and the X-fusion detects and localise the right silhouettes. Other challenging situations present in ONERA.ROOM are shown in Figures 8 (people in bright, sun-illuminated environments), Figure 9 (people crouching or lying on the ground) and Figure 10 (multiple people occluding each other). These situations are examples of the X-fusion strength versus single experts, explaining the gain of performance shown in Table III.

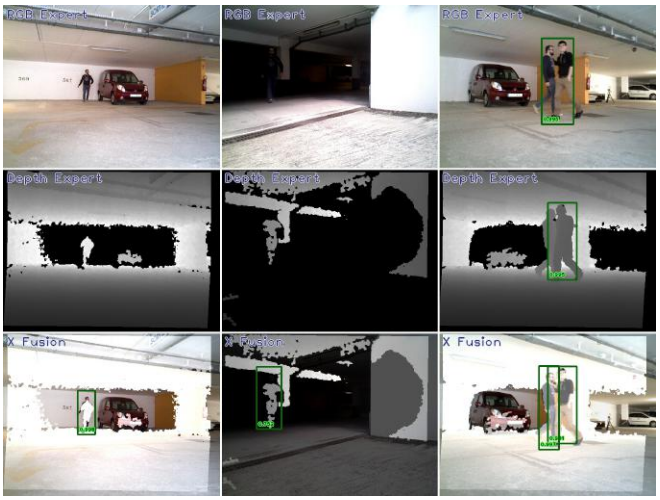

Fig. 8. X-fusion on ONERA.ROOM allows to make a detection where both RGB/D experts were impotent (left and middle column) and can differentiate two very close people mingled by both RGB/D experts (right column).

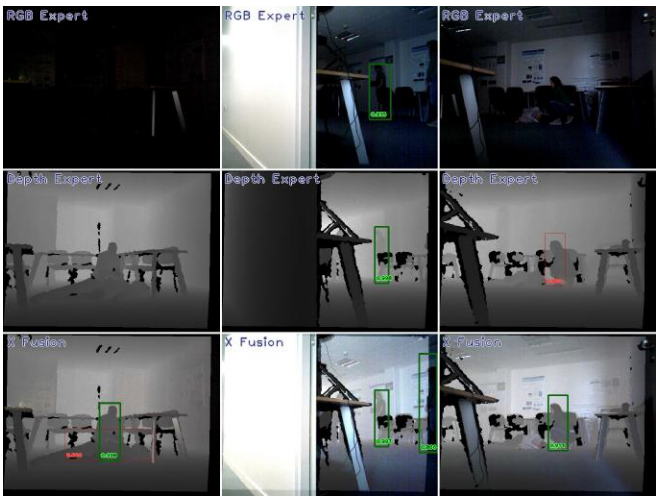

Fig. 9. X-fusion on ONERA.ROOM was close to detect the unconscious person on the floor (left column) but lacks of precision in the ROI. Still, it is the only one to detect a crouching person (left column), a person at the edge of the depth rectified image (middle column), and propose a better ROI in the last column than the depth expert.

IV. CONCLUSION

We have presented several strategies for merging the predictions of CNNs experts on different modalities. The multimodal object detection architecture based on *Faster RCNN* [1] enhances robustness in the case of heterogeneous conditions. The results on the InOutDoor RGBD People [2], RGBD People [3] and ONERA.ROOM datasets show that these strategies result in an average precision gain under normal conditions and remain robust under extreme conditions. In addition, we set new references in the state of the art on these datasets. Our best proposal, the X RGBD RCNN, gets more than 90% of AP on all these datasets and is able to withstand the failure of one of the two sensors. Lastly, we made ONERA.ROOM publicly available in the hope that it will encourage and facilitate the work on challenging RGBD

data.

V. PERSPECTIVE AND FUTURE WORK

Our future work will aim at incorporating temporal information to enable re-identification of previous detections and to implement temporal filtering of class probability vectors. A detection tracker could be seen as a third expert proposing ROIs previously revealed: "Look at this one, again!". As mentioned in [13], the NMS here is still a hand crafted processing who can be improved by deep learning. This is particularly interesting considering that both experts here are equally weighted whereas the RGB expert alone is better than the depth expert. Thus, a trained NMS could benefit from this kind of *a-priori* knowledge.

VI. ACKNOWLEDGEMENTS

We would like to thank all of the ONERA.ROOM participants and the ONERA "ATEXPA TEAM", especially Martial Sanfourche, Anthelme Bernard-Brunel, Aurélien Plyer and Hélène Roggeman for their contribution to ONERA.ROOM.

Fig. 10. X-fusion on ONERA.ROOM allows to improve ROIs proposition (left and middle columns) and is even able to detect the unconscious person (right column). However, as for the U-fusion, these strategies suffer from false positive (left column) because they can not remove a detection.

REFERENCES

[1] 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.

[2] O. Mees, A. Eitel, and W. Burgard, "Choosing smartly: Adaptive multimodal fusion for object detection in changing environments," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016.

[3] L. Spinello and K. O. Arras, "People detection in rgb-d data," in 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2011.

[4] 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, pp. 886–893, IEEE, 2005.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010.

[6] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in European Conference on Computer Vision, 2014.

[7] 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.

[8] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015.

[9] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in Proc. ACCV, vol. 2, 2016.

[10] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," arXiv preprint arXiv:1511.00561, 2015.

[11] B. Steder, G. Grisetti, M. Van Loock, and W. Burgard, "Robust on-line model-based object detection from range images," in Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on, pp. 4739–4744, IEEE, 2009.

[12] R. Girshick, "Fast R-CNN," in Proceedings of the IEEE International Conference on Computer Vision, 2015.

[13] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," arXiv preprint arXiv:1705.02950, 2017.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv preprint arXiv:1512.03385, 2015.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 248–255, IEEE, 2009.

[17] M. Luber, L. Spinello, and K. O. Arras, "People tracking in RGBD data with on-line boosted target models," in Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on, pp. 3844–3849, IEEE, 2011.

[18] L. Spinello and K. O. Arras, "Leveraging RGBD data: Adaptive fusion and domain adaptation for object detection," in Robotics and Automation (ICRA), 2012 IEEE International Conference on, pp. 4469–4474, IEEE, 2012.