Choosing Smartly: Adaptive Multimodal Fusion for Object Detection in Changing Environments

Oier Mees* Andreas Eitel* Wolfram Burgard

Abstract— **Object detection is an essential task for autonomous robots operating in dynamic and changing environments. A robot should be able to detect objects in the presence of sensor noise that can be induced by changing lighting conditions for cameras and false depth readings for range sensors, especially RGB-D cameras. To tackle these challenges, we propose a novel adaptive fusion approach for object detection that learns weighting the predictions of different sensor modalities in an online manner. Our approach is based on a mixture of convolutional neural network (CNN) experts and incorporates multiple modalities including appearance, depth and motion. We test our method in extensive robot experiments, in which we detect people in a combined indoor and outdoor scenario from RGB-D data, and we demonstrate that our method can adapt to harsh lighting changes and severe camera motion blur. Furthermore, we present a new RGB-D dataset for people detection in mixed in- and outdoor environments, recorded with a mobile robot. Code, pretrained models and dataset are available at `http://adaptivefusion.cs.uni-freiburg.de`**

I. INTRODUCTION

Most autonomous robots operating in complex environments are equipped with different sensors to perceive their surroundings. To make use of the entire sensor information in the context of an object detection task, the perception system needs to adaptively fuse the raw data of the different sensor modalities. This sensor fusion is challenging for object detection because the sensor noise typically depends substantially on the conditions of environment, which even might be changing. An example scenario for a changing environment would be a robot equipped with an RGB-D sensor, that has to operate in both, a dark indoor and a bright outdoor scenario, during different times of the day, or under different weather conditions in the case of autonomous cars. Our goal is to equip the perception system of robots with the capability to autonomously adapt to the current conditions without human intervention. For example, in a dark indoor scenario, one would expect the depth information from an RGB-D device to be more reliable than the visual appearance. Additionally, the depth stream would not be very informative in a sunny outdoor scenario with objects that are far away from the robot. In this paper, we demonstrate that such prior information can be learned from raw data and without any hand-crafted features. Therefore, how to best combine different modalities for robust object detection is the main question we tackle. We make the following contributions:

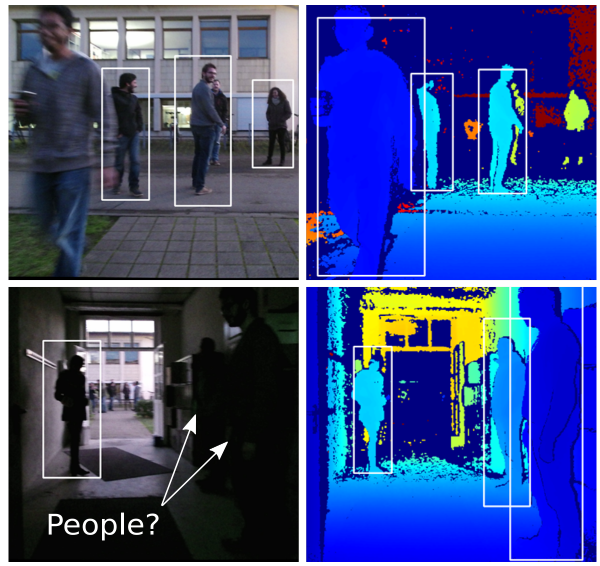

Fig. 1: Vision-based detection fails in very dark environments and under motion blur, while it performs good in outdoor scenarios and for larger distances, where depth images are usually noisy. Our approach combines the best of both worlds in an adaptive fusion manner.

- We introduce a novel fusion scheme for object detection, based on a mixture of deep network experts.
- We learn the adaptive fusion using a CNN, that is trained to weight the expert classifier outputs, based on high-level features extracted from the expert networks, without the use of prior information.
- We evaluate our method in extensive real-world experiments and demonstrate that it is more robust in changing environments than purely vision-based or other multimodal fusion approaches.

Although our method is applicable for an arbitrary number of object classes, in this work we report results for the two-class problem of human detection in the RGB-D domain. First, our experimental results show an increased performance of our method compared to other fusion approaches reported on the publicly available RGB-D People Unihall dataset [18]. Second, we evaluate our approach on a more challenging detection scenario. We recorded RGB-D sequences of people from a mobile robot, captured under abrupt changes in lighting conditions, both indoors and outdoors. In comparison to previously recorded datasets for mobile robots, the sequences

*These authors contributed equally.

All authors are with the Department of Computer Science, University of Freiburg, Germany. {meeso, eitel, burgard}@informatik.uni-freiburg.de.
This work has partly been supported by the European Commission under ERC-AGPE7-267686-LIFENAV.

show the robot moving through poorly illuminated indoor environments, followed by very bright outdoor scenes within a short period of time. Throughout this paper, when talking about a changing or a dynamic environment, we mean that the underlying conditions of the environment are changing. Examples of underlying conditions are

- lighting changes,
- out of range readings in the depth sensor,
- reflective materials that cause sensor noise in the depth channels of an RGB-D sensor and
- motion blur.

II. RELATED WORK

Sensor fusion for robust object detection in changing environments is at the core of many robotics applications. In the past, a large fraction of work has focused on this task in the context of human and also general object detection. Yebes *et al.* [24] combined features from appearance and depth data in conjunction with a DPM detector to recognize objects in road scenes. The features are merged in a channel fusion manner, by modelling a 3D-aware HOG descriptor. In comparison to their method, our approach aims to combine features at a later stage. Enzweiler *et al.* [6] introduced a mixture of experts approach for people detection using three input modalities, namely appearance, motion and depth. In comparison to our approach the weighting of the experts is constant and therefore not adaptive. Premebida *et al.* [16] trained a late fusion SVM with manually designed features to perform detector fusion from RGB and depth modalities. In our work, we do not use any prior information to learn the weights for the fusion, although it would be possible to use our method with manually designed fusion features. Spinello *et al.* [19] proposed a hierarchical mixture of experts approach, where the output of the individual detectors is weighted based on missing information in the sensor modalities. Compared to their approach, our weighting function is directly learned from a feature representation of the raw input data. Very recently, late-fusion network architectures have shown to be very successful for vision-tasks such as multimodal pedestrian detection [17], [22], RGB-D object recognition [5] and RGB-D object detection [8]. Therefore, we will test similar late-fusion network architectures as baselines for comparison with our approach.

Our work is also related to the area of multimodal people detection, with emphasis on mobile platforms equipped with RGB-D sensors. Jafari *et al.* [10] proposed a multimodal approach which combines a depth-based close distance upper body detector and an appearance-based full body detector based on groundHOG features. Detections of both modalities are fed into a multi-hypothesis EKF-based tracking module. In comparison to filtering over time, our approach performs fusion on a per-frame basis. For further comparisons, we refer the reader to the work of Spinello *et al.* [18], Munaro *et al.* [14] and Linder *et al.* [13].

There has been a large body of research targeting pedestrian detection in the vision community [21], [15]. For a concise overview of pedestrian detection we refer to a recent discussion by Benenson *et al.* [2]. They conclude that improved detection performance has shown to be driven by the design of better features, but also complemented by additional data, such as image context and motion. To demonstrate the quality of features learned by a convolutional neural network, Hosang *et al.* [9] reported improved results for pedestrian detection, using pre-trained off-the-shelf CNNs. More recently Angelova *et al.* [1] presented a convolutional network architecture that processes larger areas of the RGB input image and detects multiple pedestrians simultaneously, leading to a significant speedup at test time. However, none of these approaches make use of multiple modalities.

III. MIXTURE OF DEEP NETWORKS ARCHITECTURE

Our detection approach is based on a mixture of deep network experts (MoDE) that are fused in an additional network, which we will further denote as gating network. The overall architecture, which is illustrated in Fig. 2, is an extension of the adaptive mixture of experts method [11] that differs in the modelling of the gating network. It takes as input the feature representations extracted from a higher level in the hierarchy of each expert network, instead of using the raw pixel input as presented in the original method. The gating network then decides, based on its input, how to weight the outputs of each expert to produce the final classifier output. Let $\mathcal{D} = \{(\mathbf{x}^1, \mathbf{y}^1), \dots, (\mathbf{x}^N, \mathbf{y}^N)\}$ be the training examples, where $\mathbf{x}^k = (x_1, \dots, x_M)$ denotes a sequence of matrices which describe M different input modalities along with a gating network g. The output encoding the class labels is defined as a C dimensional vector $\mathbf{y} \in \mathbb{R}^C$ in one-hot encoding.

A. Fusion via mixture of experts

We combine the classifier outputs $f_i(x_i) \in \mathbb{R}^C$ of $i = 1, \dots, M$ experts – one for each modality – by gating functions $g_i(r(\mathbf{x})) \in [0, 1]$ with $\sum_{i=1}^{M} g_i(r(\mathbf{x})) = 1$. Let h denote a feature map produced by the last pooling layer of each expert CNN. It can be described as a three-dimensional array of size $Nh \times Hh \times Wh$ (number of filters, height, width). The resulting concatenated feature map h_{concat} for M experts is an array of size $(Nh \cdot M) \times Hh \times Wh$. The flattened representation of h_{concat} is denoted as $r(\mathbf{x})$ which is a one-dimensional array of size $1 \times (Nh \cdot M \cdot Hh \cdot Wh)$.

Therefore the gating functions depend on the input $\mathbf{x}$ solely through the representation $r(\mathbf{x})$, which significantly minimizes the input dimension of the gating network. Each $f_i(x_i)$ maps the raw pixel input to the C outputs $c = 1, \dots, C$ and the fused classifier output can then be written as

$$F(\mathbf{x}) = [F_c(\mathbf{x})]_{c=1}^{C} = \sum_{i=1}^{M} g_i(r(\mathbf{x})) f_i(x_i). \qquad (1)$$

This can be formulated as probability model so that

$$F_c(\mathbf{x}) \simeq \sum_{i=1}^{M} p(e_i \mid r(\mathbf{x})) \cdot p(c \mid e_i, x_i) = p(c \mid \mathbf{x}), \qquad (2)$$

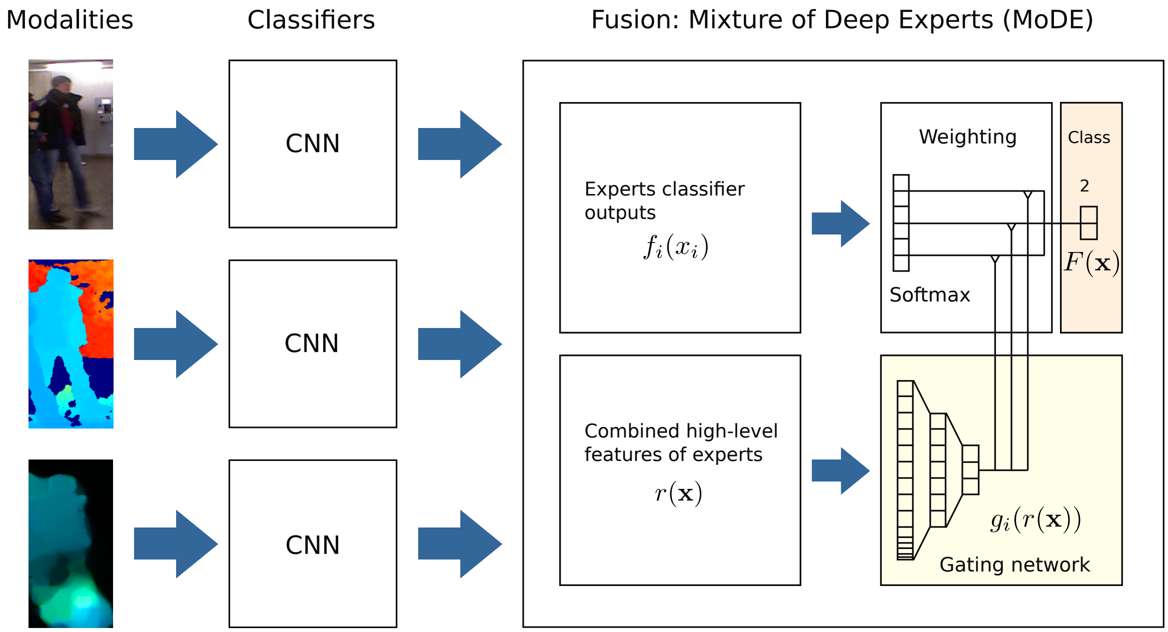

Fig. 2: We propose a mixture of deep network experts for sensor fusion, which improves classification of several CNN experts trained on different input modalities through an additional gating network. The fusion is learned in a supervised way, where the gating network is trained with a feature representation of the expert CNNs. For each input sample, the gating network assigns an adaptive weighting value to each expert to produce the final classification output.

where $p(e_i \mid r(\mathbf{x}))$ denotes the probability of selecting an expert i and $g_i(r(\mathbf{x})) = p(e_i \mid r(\mathbf{x})) = \text{softmax}(\xi_i)$ represents the contribution of a single expert to the final probability over classes $p(c \mid \mathbf{x})$. The i^{th} output of the last inner product layer of the gating network is denoted as ξ_i and the softmax function is defined as $\text{softmax}(z) = \exp(z)/\sum_j \exp(z_j)$. We train the combined gating network architecture using the cross-entropy loss that is defined as:

$$L(\mathbf{w}) = -\frac{1}{N}\sum_{k=1}^{N} \mathbf{y}^{k^T} \log F(\mathbf{x}^k). \tag{3}$$

The final architecture of our model is depicted in Fig. 2. For the following detection tasks we define $C = 2$, representing the two classes background and human.

B. Training a mixture of experts

We train our model using a two stage approach, where in the first stage we train the individual expert networks in an end-to-end manner using stochastic gradient descent (SGD). In this paper, we use several expert architectures, a standard three-layer convolutional neural network as a baseline and a more evolved deep network based on the Google inception architecture [20]. The three-layer convolutional neural network is a small network designed to solve the CIFAR-10 classification problem and will be further depicted as CifarNet. It has also been proposed by Hosang *et al.* [9] as a good baseline for people detection. In Section IV-B we consider a downsampled Google inception architecture, depicted as GoogLeNet-xxs. For our first experiment in Section IV-A, the CifarNet baseline already outperforms previous reported approaches. In order to leave room for improvement via sensor fusion, we report results with the CifarNet as architecture for the single experts and the mixture of experts network. In our second experiment, we show that further improvements over the CifarNet baseline can be achieved, by replacing the network with an inception architecture. For the GoogLeNet-xxs network, we only use the layers of the original network up to the first softmax classifier ("softmax0"). For training the networks, we use Fast R-CNN [7], including the region of interest pooling layer and the multi-task loss for bounding box regression. The framework is wrapped around a modified version of the Caffe library [12]. All experts are trained using standard parameters and in the first stage we apply dropout, as a regularizer, only in the fully-connected layers. Further, the networks of the RGB modality are fine-tuned using the pre-trained models available from the Caffe library [12], whereas the experts for the other domains are trained from scratch. In the second training phase, the gating network is trained on an additional validation set. We optimize the weights of the gating network using SGD and keep the weights of the individual experts fixed, by setting the learning rate of all layers to zero. Although the weights of the experts are not changed, we apply a modification to the expert layers for training the gating network. In all expert layers we now apply dropout, as a special case of data augmentation, in order to improve the performance of the gating network. To generate region of interest proposals for the networks, we implemented a dense multiscale sliding window approach. Throughout the paper we use different input modalities to feedforward through the networks. We mainly use a combination from the following modalities: RGB, depth and optical flow (for representing motion). For optical flow computation we use the OpenCV

Input	Method	IoU	AP/Recall	EER
Depth	CifarNet	0.6	78.0/88.0	84.5
RGB	CifarNet	0.6	66.5/73.0	71.8
OF	CifarNet	0.6	52.9/58.2	58.0
D-OF	CifarNet average	0.6	69.3/34.2	78.6
D-OF	CifarNet late fusion	0.6	78.6/86.0	83.7
D-OF	CifarNet MoDE	0.6	78.1/88.5	86.0
RGB-D-OF	CifarNet average	0.6	77.0/83.8	79.2
RGB-D-OF	CifarNet late fusion	0.6	88.0/88.4	88.2
RGB-D-OF	CifarNet MoDE	0.6	**88.6/90.0**	**89.3**

TABLE I: Performance of different single and multimodal networks on the RGB-D People dataset.

Input	Method	IoU	EER
Depth	HOD + Segm [4]	0.4	84.0
Depth	HOD + Sliding Window [18]	0.4	83.0
Depth	HOD [19]	0.6	56.3
RGB-D	Combo-HOD [18]	0.4	85.0
RGB-D	HGE* [19]	0.6	87.4
Depth	Ours, CifarNet	0.4	**93.7**
RGB-D-OF	Ours, CifarNet MoDE	0.6	**89.3**

TABLE II: Comparison of our best performing models with other approaches reported on the RGB-D People dataset.

implementation of the method proposed by Brox *et al.* [3].

IV. Experiments

A. RGB-D People Unihall Dataset

We evaluate our approach on the publicly available RGB-D People Unihall dataset provided by Spinello *et al.* [18]. The dataset contains over 3,000 frames of people passing by three vertically mounted Kinect cameras in a university hallway. As evaluation metrics, we compute average precision (AP) and equal error rate (EER). We define the equal error rate as the point in the precision-recall curve where precision and recall values are equal. Adopting their no-reward-no-penalty policy, we do not count true positives or false positives when a detection matches an annotation of a partially occluded person. For training, we randomly select 700 frames from each of the three Kinect cameras and extract positive samples from the annotated candidates that show fully visible people. For evaluation, we use the remaining 300 frames from each of the three Kinects. To set hyperparameters of the learning procedure we evaluate all trained models on the training set, choosing the best performing model for evaluation on the test set. Moreover, the dataset does not provide a pre-defined train/test split. Therefore we created five random train/test splits, to train and evaluate our detector. We obtained a standard deviation of ±0.8 EER, showing a small influence of the splits chosen. For some experiments we report results using both an intersection over union (IoU) of 0.4 and 0.6, due to different evaluation metrics used in the literature. Unless otherwise specified, we use an IoU of 0.6 for evaluation, which is the area of the overlap between a predicted bounding boxes and an annotated ground truth box.

1) Single expert performance We compare all reported approaches on the RGB-D People dataset that use HOD features against our baseline architecture (CifarNet) trained on colorized depth data. We further report performance for the three single networks, trained on three input modalities, namely RGB, depth and optical flow. The depth images are pre-processed using the jet colorization methodology presented in our own previous work [5]. Table I shows that the CNN trained on depth data (Depth CifarNet) performs better in comparison to the network trained on the RGB modality (RGB CifarNet), which matches previous results on the dataset [18]. We hypothesize the reason for this to be due to illumination changes and low contrast found on the RGB images. Besides, these lead to a noticeable amount of corruption when computing the optical flow images. Consequently, the optical flow CNN underperforms when compared with the other experts, also because it is limited to detect moving people only. We show in Table II, that our Depth CifarNet outperforms HOD based approaches. When evaluating the detection at IoU 0.4, our best network achieves an EER of 93.7%, which is a relative improvement of 9.7% over the graph based segmentation combination with HOD proposed by Choi *et al.* [4]. We report an improvement of 28.2% over previous approaches when evaluating the same model at IoU 0.6.

2) Comparison of fusion approaches We compare detection performance of different fusion approaches when combining 1) depth and motion (D-OF) and 2) depth, RGB and motion (RGB-D-OF). A naive way of fusing the individual detectors is averaging the classifier outputs $F(\mathbf{x}) = \sum_{i=1}^{M} \frac{1}{M} f_i(\mathbf{x}_i)$. As can be seen in Table I, this approach looses AP and EER points with respect to other methods. A more involved method is a late fusion approach, where a fully-connected layer is trained on top of the last pooling layers of all network streams. Formally, we concatenate the last layer feature responses $r(\mathbf{x})$ and feed them through an additional two-layer fully-connected fusion stream $f(r(\mathbf{x}); \theta)$ with parameters θ. The first fully-connected layer has 64 outputs and is followed by a second layer of output two, that ends in a softmax classifier. We follow the same two stage training procedure described earlier. The results given in Table I show, that the late fusion approach substantially improves detector performance when combining all three modalities. A combined depth and optical flow late fusion approach underperforms and we conclude that it is not suitable for modalities that highly differ in their feature space. The architecture of the gating network for the proposed MoDE consists of two fully-connected layers with rectified linear units of size 64 and three. The input of the gating network is the combined output of the last pooling layer of each expert. The proposed MoDE approach yields best performance for all fusion scenarios. When combining all three modalities (MoDE) we achieve an EER of 89.3%. In addition, we report a relative improvement of 1.9% EER over HGE's 87.4% EER [19].

B. InOutDoor RGB-D People Dataset

The next experiment is conducted in a challenging people detection scenario recorded from an Kinect v2 camera, that was mounted on a mobile robot. Several abrupt and harsh lighting changes, combined with severe motion blur provide

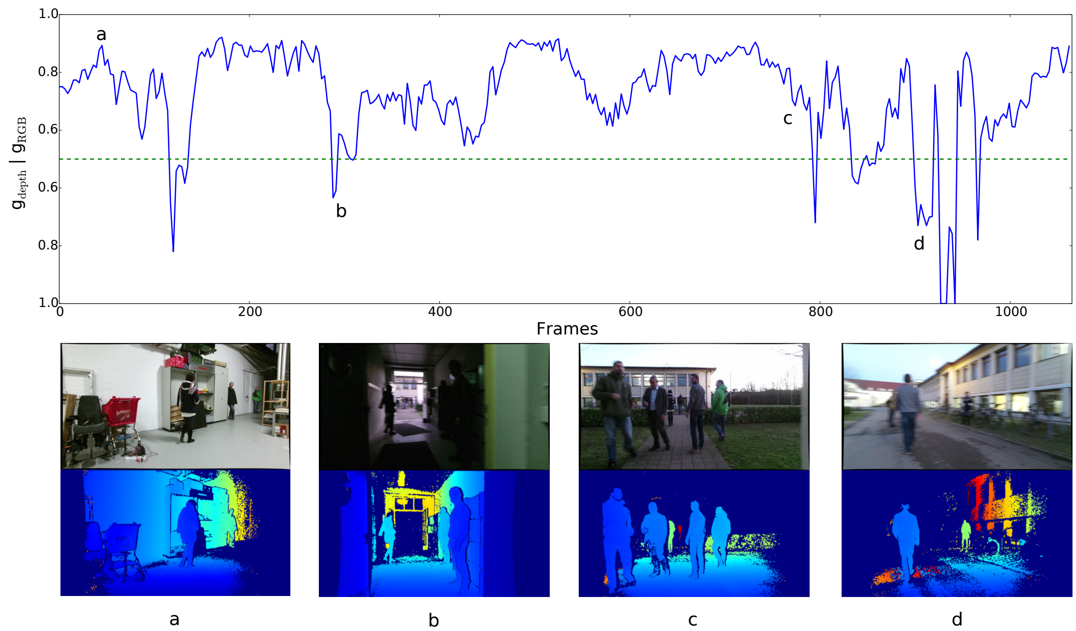

Fig. 3: Timeline of the gating weights on the test sequence. The gating assignment weight switches between the RGB and depth modalities depending on the current environment (a, b, c, d). For example at frame b, the people are hard to detect in RGB, because the scene is dark. Accordingly, the assigned gating RGB weight is low $g_{\text{RGB}} = 0.38$. The respective depth expert's gating weight is higher $g_{\text{depth}} = 1 - g_{\text{RGB}} = 0.62$, due to people being better visible in the depth image.

challenges to both sensor modalities RGB and depth, see Fig. 1. A particularity of the recorded sequences is, that the robot is driving from indoor to outdoor environments in a single take. We recorded and annotated a total of 8605 RGB-D frames collected from the robot at a frame rate of 30hz. The camera was calibrated using the approach of Wiedemeyer *et al.* [23].

The dataset is subdivided into four sequences, two recorded during midday and the other two recorded at dusk. The test sequence contains 1,066 frames, during which the robot drives from a dark indoor scene into a relatively dark street scene, see Fig. 3. As evaluation metrics, we compute average precision (AP) and use an IoU of 0.6 for evaluation of positive detections.

1) Quantitative Results We report performance of several single modality networks, namely RGB and depth and show improved results when fusing the modalities, see Table III. Further, we compare the CifarNet architecture with the more evolved GoogLeNet-xxs architecture. All single modality networks perform reasonably well on the test set. Our depth-based GoogLetNet-xxs network outperforms the upper body detector of Jafari *et al.* [10]. The gating network architecture for the GoogLeNet-xxs consists of a convolution layer with output size 128 and kernel size 3, followed by two fully-connected layers of size 128 and 2. The combined outputs of the last convolutional layers of the expert networks serve as input to the gating network. Our novel fusion approach improves the performance by 8.4% AP in comparison to the late fusion approach. The best model is a combination of the RGB and depth GoogLeNet-xxs, fused with the gating network (GoogLeNet-MoDE). We also evaluated switching to the expert that the gating network predicted, defining $g_i(r(\mathbf{x})) \in \{0, 1\}$ as in [11], instead of weighting the experts, but found it to underperform. For further comparison, we trained a multi-channel network with six channels as input in an end-to-end manner from scratch. The network slightly underperforms in comparison to the late fusion approach using the same GoogLeNet architecture. The MoDE method also outperforms a naive averaging of the classifier outputs.

Input	Method	AP/Recall
RGB	GoogLeNet-xxs	70.0/79.6
RGB	CifarNet	55.3/62.9
Depth	GoogLeNet-xxs	71.6/78.9
Depth	Upper-body detector [10]	69.1/72.0
RGB-D	GoogLeNet average	71.1/73.9
RGB-D	GoogLeNet channel fusion	71.0/79.4
RGB-D	GoogLeNet switching [11]	71.0/73.9
RGB-D	GoogLeNet late fusion	72.0/76.3
RGB-D	GoogLeNet-MoDE	**80.4/81.1**

TABLE III: Performance of single and multimodal networks on the InOutDoor RGB-D People dataset.

2) Qualitative results When analyzing the output detections obtained with the RGB-D based MoDE detector over the full test sequence, we gain several interesting insights. Fig. 3 shows for each frame the average of the gating weights for both experts. The plot shows that the average gating weights correlate reasonably well with the respective environment. In the bright indoor scenario (sequence a), the RGB modality is chosen more often, whereas in dark environments (sequence b) the gating network tends to weight the depth network output higher than the vision-based one. Outdoors (sequence c), it mainly relies on RGB, especially for the pedestrian at far distance who is almost not visible in the depth image, as can be seen in Fig. 4. We

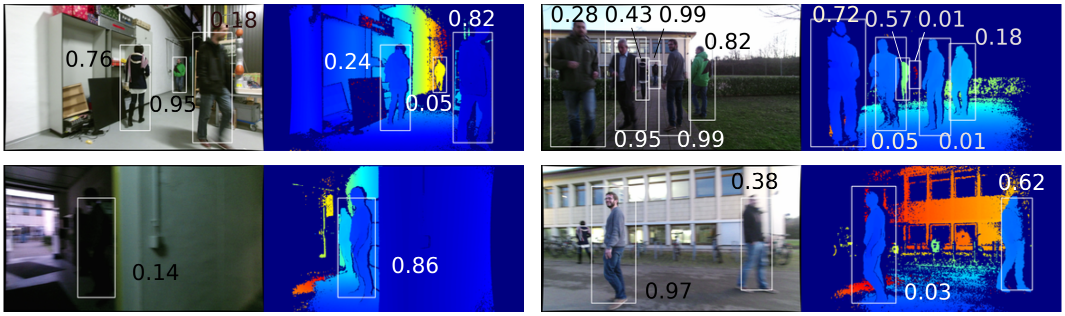

Fig. 4: Qualitative results of people detection for the RGB-D based MoDE detector. We show the assignment weights $g_i(r(\mathbf{x}))$ learned by the gating network for each expert.

also found the depth network to perform better for blurred images (sequence d). Therefore, in the final frames of the test sequence, the gating weights vary more in between the frames. Here, the RGB outputs are mostly chosen, although for abrupt camera motion the gating network switches with higher frequency. Fig. 4 shows the gating weights for several exemplary frames, supporting our observation that the gating network tends to switch for changing underlying conditions such as lighting conditions, false depth readings at far ranges and motion blur.

V. Conclusion

In this paper, we considered the problem of multimodal adaptive sensor fusion for object detection in changing environments. We proposed a novel mixture of deep network experts approach that automatically learns an adaptive strategy for weighting several domain-based classifiers, from the raw input data. In extensive experiments, we demonstrated that our multimodal method outperforms a vision-based detection baseline and other fusion techniques. Moreover, we show improved detection performance in a sequence recorded from a mobile robot containing abrupt lighting changes and severe motion blur. Finally, our system outperforms previously reported approaches for depth-based people detection on the publicly available RGB-D People and our InOutDoor People dataset.

Acknowledgments

We would like to thank Jost Tobias Springenberg, Luciano Spinello and Gian Diego Tipaldi for helpful discussions.

References

[1] A. Angelova, A. Krizhevsky, and V. Vanhoucke, "Pedestrian detection with a large-field-of-view deep network," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2015.

[2] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *ECCV workshop on computer vision for road scene understanding and autonomous driving*, 2014.

[3] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *European Conference on Computer Vision (ECCV)*, 2004.

[4] B. Choi, C. Mericli, J. Biswas, and M. Veloso, "Fast human detection for indoor mobile robots using depth images," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2013.

[5] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015.

[6] M. Enzweiler and D. M. Gavrila, "A multilevel mixture-of-experts framework for pedestrian classification," *IEEE Transactions on Image Processing*, 2011.

[7] R. Girshick, "Fast R-CNN," in *Int. Conf. on Computer Vision (ICCV)*, 2015.

[8] 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 (ECCV)*, 2014.

[9] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[10] O. Hosseini Jafari, D. Mitzel, and B. Leibe, "Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2014.

[11] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, 1991.

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. of the ACM Int. Conf. on Multimedia*, 2014.

[13] T. Linder, F. Girrbach, and K. O. Arras, "Towards a robust people tracking framework for service robots in crowded, dynamic environments," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015.

[14] M. Munaro and E. Menegatti, "Fast RGB-D people tracking for service robots," *Autonomous Robots*, 2014.

[15] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[16] C. Premebida, J. Carreira, J. Batista, and U. Nunes, "Pedestrian detection combining rgb and dense lidar data," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2014.

[17] J. Schlosser, C. K. Chow, and Z. Kira, "Fusing lidar and images for pedestrian detection using convolutional neural networks," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2016.

[18] L. Spinello and K. O. Arras, "People detection in RGB-D data," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2011.

[19] ——, "Leveraging RGB-D data: Adaptive fusion and domain adaptation for object detection," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2012.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[21] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *European Symp. on Artificial Neural Networks (ESANN)*, 2016.

[23] T. Wiedemeyer, "IAI Kinect2," www.github.com/code-iai/iai_kinect2, 2014 – 2015.

[24] J. J. Yebes, L. M. Bergasa, and M. García-Garrido, "Visual object recognition with 3d-aware features in kitti urban scenes," *Sensors*, 2015.