2012 IEEE International Conference on Robotics and Automation
RiverCentre, Saint Paul, Minnesota, USA
May 14-18, 2012

Leveraging RGB-D Data: Adaptive Fusion and Domain Adaptation for Object Detection

Luciano Spinello and Kai O. Arras

Social Robotics Lab, University of Freiburg, Germany
{spinello, luber, arras}@informatik.uni-freiburg.de

***Abstract*— Vision and range sensing belong to the richest sensory modalities for perception in robotics and related fields. This paper addresses the problem of how to best combine image and range data for the task of object detection. In particular, we propose a novel adaptive fusion approach, hierarchical Gaussian Process mixtures of experts, able to account for missing information and cross-cue data consistency. The hierarchy is a two-tier architecture that for each modality, each frame and each detection computes a weight function using Gaussian Processes that reflects the confidence of the respective information. We further propose a method called cross-cue domain adaptation that makes use of large image data sets to improve the depth-based object detector for which only few training samples exist. In the experiments that include a comparison with alternative sensor fusion schemes, we demonstrate the viability of the proposed methods and achieve significant improvements in classification accuracy.**

I. Introduction

Object detection is fundamental to robotics, computer vision, human-computer interaction, and intelligent vehicles. RGB-D sensors are particularly interesting for this task as they provide both range and image data. Thus, the question arises how to leverage RGB-D data and optimally combine the sensory modalities so as to achieve efficient learning and particularly robust object detection – the topic addressed in this paper.

What makes such a sensor fusion task a challenge is, for instance, varying amounts of missing data in one or both of the modalities. When this occurs, the fusion architecture should gracefully degrade and still provide a meaningful result over wide ranges of invalid inputs. The architecture should also exploit the complementarity of the data by detecting contradictions and a way to adaptively compute weighting for such situations. For the important task of object detection, a question arises if large data sets existing only in one modality can be used to learn a robust detector in another modality. Here, we address these points in the context of object detection in image and range data.

In robotics, range sensors have been popular for their accuracy, robustness against illumination changes and vibration, large field of view and calibration-free employment. Early works in 2D object detection are typically concerned with dynamic objects in traffic scenes such as people using model-based or learned classifiers [5, 18]. Object detection in 3D range data continues to be an active area of research, recent contributions in a robotics context include [15, 30, 28, 29]. Cameras provide rich and dense information of the scene which is why visual object detection has been extensively studied in the field of computer vision. Popular approaches include [4, 16, 12, 10, 6].

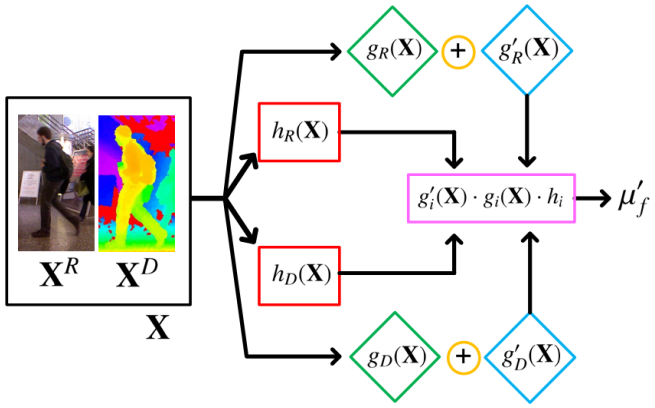

Fig. 1. Overview of the multi-modal fusion architecture for RGB-D data based on a hierarchy of Gaussian Process mixtures of experts.

Several researchers have combined image and range data for object detection: [7, 27] propose a trainable 2D range data and camera detection system for multiple object categories. [31, 26] combine multiple modalities such as dense depth data, optical flow maps, and images to detect people in urban environments. None of the cited works adaptively fuse information. The fusion schemes are fixed and do not change with the properties of the data. A form of adaptation is proposed in [9, 8] that modify their fusion scheme based on the detection of local occlusions or object shape fitting. The approach proposed here is more general as we will detail further below.

The other aspect addressed in this paper is efficient learning from multi-modal data. The large majority of object detectors today are based on supervised learning. Clearly, the achievable accuracy of such a detector heavily depends on the size and quality of the available training set. Hundreds, thousands or more instances of the object in different poses, illuminations, and in front of different backgrounds are required for the creation of a robust classifier. The collection and annotation of such data is time-consuming and expensive. Approaches to generate data from synthetic images exist but require hard-to-design photorealistic 3D models of the object category [22, 15, 24].

To address these issues, we make the following contribu-

978-1-4673-1405-3/12/$31.00 ©2012 IEEE

tions in this paper:

- We introduce an adaptive fusion method of image and depth data for object detection based on a hierarchical mixture of experts approach. The technique computes a weight for each modality that accounts for missing information and cross-cue data consistency. To our knowledge, the proposed fusion scheme is the first to assume independency between the modalities able to work over all combinations of missing information from one or both cues. This generalizes the fusion approaches in [9, 8] where depth data are gained from stereo and thus depend on the image data.
- The approach fuses information adaptively, that is different on each frame and location in the image. The fusion weights are dynamic and change as a function of the multi-modal data conditions in different parts of the image.
- We employ Gaussian Processes as mixing components of the layers in the fusion architecture. With this method we are able to interpolate probabilistically among sparse (non-uniformly distributed) and scarce (low in number) training samples.
- We introduce *cross-cue domain adaptation*, a method that allows to use large RGB data sets to train a depth-based detector for which few training samples exist.

We evaluate the methods using people as object category, one of the most challenging objects due to the variability of human shape and appearance. We consider several alternative fusion techniques and show the beneficial effects of cross-cue domain adaptation onto the classification performance.

The paper is structured as follows: the next section presents the detection methods used in the paper, followed by Section III that describes the proposed fusion architecture. Section IV introduces the cross-cue domain adaptation technique. Section V contains details on the experiments and the results. Section VI concludes the paper.

For the sake of notation simplicity, we will denote all combinations of image and range data as *RGB-D data* in the remainder of the paper. Examples may include data from dense stereo, lidars with point intensity (or color) and data from structured light sensors such as Microsoft Kinect.

II. Object Detectors for RGB-D Data

The proposed methods are general with respect to the choice of the object detector. Any window-scrolling approach can be plugged into the adaptive fusion architecture and any gradient-based detector is suited for cross-cue domain adaptation. Most visual object detectors today are based on window-scrolling [4, 12, 6, 9, 26, 25, 32]. Window-scrolling has also been applied to depth data or other modalities for object detection [31, 26, 8]. Further, a large number of successful techniques for feature extraction and recognition are based on gradient information. Examples include SIFT [17], SURF [1], HOG [4], or Centrist [32].

Specifically, we employ the recently proposed Combo-HOD detector [26] which is a multi-modal RGB-D object detector that combines the outputs of an Histogram of Oriented Gradients (HOG) detector for images with an Histogram of Oriented Depths (HOD) detector for dense depth data.

III. An Adaptive RGB-D Fusion Architecture

This section describes the proposed adaptive architecture and relates it to two main fusion schemes for object detection: early fusion and late fusion.

Early fusion techniques for object detection typically combine modalities in feature space, before classification. Late fusion techniques combine them at later stages, for example at the level of detections or semantic descriptions. A straightforward early fusion scheme is the concatenation of feature vectors computed in each modality. The learned classifier will then maximizes the separation between background and objects in the joint feature space whose number of dimensions is given by the sum of the feature dimensions from all modalities. Given that a single feature descriptor for an object is easily of dimension greater than several hundreds or more, early fusion schemes at feature level raise two issues related with high-dimensional spaces: many more training samples are required to avoid overfitting and extensive memory consumptions. In practice, a large training set of high-dimensional samples can quickly take several gigabytes in memory and long time to learn a classifier from.

Mostly for the reason of smaller training sets, we follow a late fusion scheme that combines data from each modality on the level of detections.

A. RGB-D Fusion via Hierarchical mixture of experts

An RGB-D data frame can be expressed as a matrix $\mathbf{X}$ composed by the RGB and depth image: $\mathbf{X} = (\mathbf{X}_R, \mathbf{X}_D)$. $\mathbf{X}_R^{i,j}$ is a grayscale intensity value associated to the pixel at position (i, j) in the image, $\mathbf{X}_D^{i,j}$ is a depth value associated to the pixel at position (i, j). The fusion scheme proposed here is inspired by the adaptive mixture of local experts (MLE) architecture [13, 14]. The architecture is a technique for combining the confidence output of several classifiers $h(\mathbf{X}) \in \mathbb{R}$ by gating functions $g(\mathbf{X}) \in \mathbb{R}$ that depend on the input $\mathbf{X}$. An MLE architecture for RGB-D data can then be written as

$$\mu_f = \sum_i g_i(\mathbf{X}) \cdot h_i(\mathbf{X}) \qquad i = \{R, D\} \tag{1}$$

where μ_f denotes the fused value, $h_D(\cdot)$ the depth-based object detector and $h_R(\cdot)$ the visual object detector. The contribution of each detector to μ_f is weighted by the gating function $g(\mathbf{X})$. Our fusion approach extends Eq. (1) by premultiplying an additional gating function:

$$\mu'_f = \sum_i g'_i(\mathbf{X}) \cdot g_i(\mathbf{X}) \cdot h_i(\mathbf{X}) \qquad i = \{R, D\}. \tag{2}$$

The extension introduces a two-tier hierarchy (see Fig. 1) in which the gating functions of the first level $g_i(\cdot)$ shape the confidence of each detector and the gating functions of the second level $g'_i(\cdot)$ adjusts the relative importance of one modality with respect to the other by setting $g'_R(\cdot) = 1 - g'_D(\cdot)$.

Without loss of generality, it is often possible to model the classifier output as a probability [3, 20]:

$$p(y \mid \mathbf{X}, \theta_i) \simeq r_i(h_i(\mathbf{X})) \qquad i = \{R, D\} \tag{3}$$

where $r_i(\cdot)$ is the function that maps the classifier output to probabilities(e.g. a sigmoid), θ_i are the parameters learned for the classifier $h_i(\cdot)$ and y indicates the existence of an object in $\mathbf{X}$. Following [14], our approach can be written probabilistically as

$$p(y \mid \mathbf{X}, \boldsymbol{\theta}) = \sum_i g'_i(\mathbf{X}) \cdot g_i(\mathbf{X}) \cdot p(y \mid \mathbf{X}, \theta_i) \qquad i = \{R, D\}. \tag{4}$$

Note that Eq. (4) encodes a different hierarchy than the one proposed in [14]. In our case the second level gating functions $g'_i(\cdot)$ weight the lower-layer output instead of an already summed up lower-level output.

B. Learning Gating Functions with Gaussian Processes

We now derive the gating functions $g_i(\cdot)$ for the task of object detection in image and range data.

Given an object-specific detection window $\mathbf{W}^{i,j,s}$ at pixel positions (i, j) and scale s of average size. Let $\rho(s)$ be a regression function that returns the expected range of an object at that scale. This regression function can be learned from data, see e.g. [26] for RGB-D data from a Kinect sensor. Then, we can find all depth readings that are compatible with $\mathbf{W}^{i,j,s}$. In other words, we extract the slice from $\mathbf{X}_D$ that contains the expected range of an object at scale-space position (i, j, s) of $\mathbf{W}$. Formally, with i, j ranging over the detection window,

$$n_{in}(s) = \sum_i \sum_j \delta\left(\rho(s), \mathbf{W}_D^{i,j,s}\right) \tag{5}$$

where $\delta\left(\rho(s), \mathbf{W}_D^{i,j,s}\right) = 1$ if $\rho(s)$ equals $\mathbf{W}_D^{i,j,s}$ within some tolerance and 0 otherwise. $n_{in}(s)$ is the number of depth measurements that are considered *in range* at scale s. The number allows to quantify a sort of cross-cue data consistency by verifying the actual presence of an object seen in other cues.

We further extract all measurements in $\mathbf{W}$ that represent missing data, e.g. out-of-range laser points or invalid depth pixels

$$n_{miss}(s) = \sum_i \sum_j \delta\left(\mathbf{W}^{i,j,s}\right) \tag{6}$$

where $\delta\left(\mathbf{W}^{i,j,s}\right) = 1$ if $\mathbf{W}^{i,j,s}$ satisfies the out-of-range or saturation condition within some tolerance and 0 otherwise.

We can now derive the gating functions from $n_{in}(s)$ and $n_{miss}(s)$ by first making them scale-independent. This is achieved by normalization with the product of the detection window area and the scale s. For notation simplicity we assume the functions to be normalized hereafter: $0 \leq n_{in} \leq 1$, $0 \leq n_{miss} \leq 1$.

The gating functions $g(\cdot)$ in the regular approach have the task to weight the confidence of their respective detector. Given n_{in} and n_{miss} for the range cue, intuitively, $g_D(\cdot)$ should be small at high values of n_{miss} and large at high values of n_{in}. Here, we learn them from training data and express them using the precision performance index. With t_p being the number of true positives and f_p the number of false positives at n_{miss}, n_{in} generated by $h(\cdot)$, we have

$$g(\mathbf{X}) = g(n_{miss}, n_{in}, \mathbf{X}_D) = \frac{t_p(n_{miss}, n_{in})}{t_p(n_{miss}, n_{in}) + f_p(n_{miss}, n_{in})} \tag{7}$$

Even though $g(\cdot)$ explains the confidence of a certain modality given the quantity of cross-cue data consistency and missing information, it does not explain the *relative importance* of a sensory cue with respect to the other. This is the purpose of the additional gating functions $g'(\cdot)$.

Intuitively, and assuming n_{miss} refers to out-of-range measurements in the range modality, $g'_R(\cdot)$ should receive high weight over $g'_D(\cdot)$ at high values of n_{miss}. Alternatively, $g'_D(\cdot)$ should receive high importance over $g'_R(\cdot)$ at high values of n_{in} which is motivated by the nature of range data that have very well defined gradients at the object's silhouette and by large values of n_{in} that strongly indicate the actual presence of the object in the image.

Again, the functions $g'(\cdot)$ are learned from a training set. They are function of the relative false negative ratio obtained by $h_R(\cdot)$ with respect to $h_D(\cdot)$ at a certain n_{miss}, n_{in}. The insight is that by reducing false positive detections with $g(\cdot)$ we are fusing two low-false-positives detectors. In this way, $g'(\cdot)$ balances the effect of $g(\cdot)$ by giving more relative importance to the modality that has statistically more probability of detecting the object.

The data necessary to learn smooth functions $g(\cdot)$ and $g'(\cdot)$ are hard to obtain: the values of the gating functions are only defined for all n_{miss}, n_{in} found in the training set. Thus, we take a Gaussian Process (GP) regression approach [21]. GPs are non-linear non-parametric models that estimate Gaussian distributions over functions based on training data. They produce a probabilistic interpolation between unevenly distributed uncertain data points. We train a GP for each gating function to learn the regression between $(n_{miss}, n_{in}) \rightarrow g(n_{miss}, n_{in})$ and $(n_{miss}, n_{in}) \rightarrow g'(n_{miss}, n_{in})$. The covariance function of GPs expresses how strong the training data are to be considered correlated. Our method makes use of a Matérn covariance function that can be estimated purely from distances between points. The two parameters of the covariance function are found by maximizing the GP marginal likelihood in the training set.

IV. RGB-D Cross-Cue Domain Adaptation

There is a good amount of established data sets for visual object detection covering various object categories ([11] or [6]). By today, this is not the case for dense 3D range data or RGB-D data. Therefore, in this section, we propose *cross-cue domain adaptation*, a method to improve a detector for which few training samples exist from annotated data of another detector.

The approach adapts features computed in the range data domain to features from the RGB data domain. This follows the insight that a local depth change is in spirit similar to a local intensity change in that both differentials describe

a shape locally, by appearance (image data) or by 3D properties (depth data). Let the image data set be the *source domain* ε_s and the range data set the *target domain* ε_t. We calculate:

- Image features computed on the three RGB channels in ε_s that describe local intensity changes (e.g. HOG [4])
- Range image features in ε_t that describe local depth changes (e.g. HOD [26])

If the depth data domain is not calibrated as is the case with RGB-D sensors, range cameras or stereo vision, we first need to relate raw depth values to range in meters. In [26] this has been done for the Kinect sensor according to its specifications in [19]. This relationship has shown to be

$$d(w) = \frac{8 \cdot B \cdot F_x \cdot D}{(W_{max} - w)} \tag{8}$$

where $B = 0.075\,m$ is the distance between the IR projector and the IR camera, F_x is the focal length of the IR camera in the horizontal direction, and $W_{max} = 1032$ is the maxium raw depth value considered in this paper. The additional term $D = 255$ maps depth values onto the RGB ranges of $\{0, 1, ..., 255\}$. In this way, d is scaled to the maximum intensity value of one RGB image channel.

Data from ε_t can now be combined with data from ε_s. A straightforward form of domain adaptation at this point is training a depth-based detector with a single combined data set from $\varepsilon_u = \varepsilon_s + \varepsilon_t$. However, as we will show in the experiments, this approach is less appropriate mainly due to the unequal training set sizes. The resulting detector will only marginally be different from a detector learned in the source domain due to quantitative dominance of its training samples in the common space.

This motivates the approach taken in [23], in which two separate classifiers $h_t(\mathbf{X})$ and $h_s(\mathbf{X})$ are trained, separately for each domain. The final domain-adapted classifier $\hat{h}(\mathbf{X})$ is then computed by a convex combination of the two real-valued classifier outputs

$$\hat{h}(\mathbf{X}) = \beta\, h_s(\mathbf{X}) + (1 - \beta)\, h_t(\mathbf{X}). \tag{9}$$

The parameter $\beta = [0, 1]$ is determined by minimizing the misclassification error on the target domain ε_t. In case of an SVM-based detectors (such as HOD or Combo-HOD), the functions $h_s(\cdot)$ and $h_t(\cdot)$ are the result of the SVM training optimization. Similar domain adaptation techniques have obtained very good results on other applications [23, 2].

V. Experiments

For the evaluation we choose people as objects for the reasons that there are well-established visual detectors trained from large data sets and that they belong to the most challenging object categories: humans are articulated objects that exhibit a large variability in their appearance from different body poses, clothing, or wearable luggage.

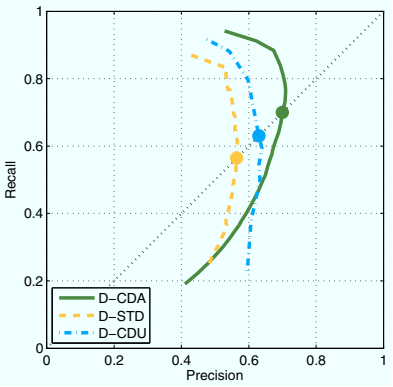

Fig. 2. Cross-cue domain adaptation approach for a range-based people detector. Precision-recall graphs for the detector that makes use of cross-cue domain adaptation (D-CDA), a simplistic domain adaptation method (D-CDU), and depth-only training data (D-STD). EER points are highlighted in the figure. D-CDA (69.9% EER) outperforms all the others thanks to the intelligent combination of the target and source domains.

A. RGB-D Data Set

We collected a large-scale indoor data set with unscripted behavior of people. The data set has been taken in the lobby of a large university canteen at lunch time. An additional data set has been collected in a visually different university building which is only used for generating background samples. The data set has been manually annotated to include bounding boxes in both domains and the visibility status of subjects (fully visible/partially occluded). A total of 1648 instances of people in 1088 frames have been labeled.

Detections are counted as true positives if the bounding box overlaps with a manually labeled person by more than 60% to account for metric inaccuracies in the annotation and the detection. Adopting the no-reward-no-penalty policy from [10], we do not count true positives or false positives when a detection matches an annotation of a partially occluded person.

B. Results on Cross-Cue Domain Adaptation

For the detection of humans in range data, we use the depth-based HOD people detector [26] (D-STD). The target domain consists of the data set from [26] containing 1030 depth data samples of people and 5000 background samples. The source domain consists of the public TUD-Brussels image data set [31], from which we used the 2184 images of people and a random set of 19200 background samples.

First we compare the domain-adapted depth-based detector, called D-CDA, with the D-STD detector, see Fig. 2. The D-CDA detector achieves a 69.9% equal error rate (EER, the point where precision equals recall), largely outperforming the baseline D-STD that has been trained with annotated range samples only (EER 56.3%).

We then evaluate the naive form of domain adaptation that learns a depth-based people detector from a single combined data set in $\varepsilon_u = \varepsilon_s + \varepsilon_t$, called D-CDU. With an EER of 63.2%, the D-CDA detector also outperforms D-CDU. The result suggests that the unified domain ε_u, that disregards

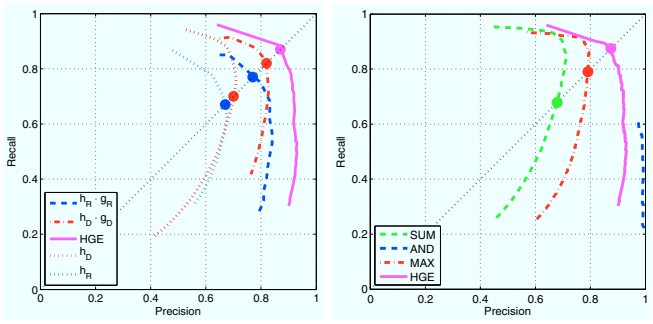

Fig. 3. Precision-recall curves for the Hierarchical Gaussian Process Mixtures of Experts approach (HGE). EER points are highlighted. **Left**: Classifier performance at places of increasing sophistication in the architecture in Fig. 1. The newly introduced gate functions $g'(\cdot)$ have a clear contribution. **Right**: Comparison of alternative sensor fusion schemes including the proposed method (HGE), maximum detector response (MAX), product of detector responses (AND) and weighted sum of detector responses (SUM). HGE outperforms all other methods thanks to the adaptivity to the data conditions in each modality.

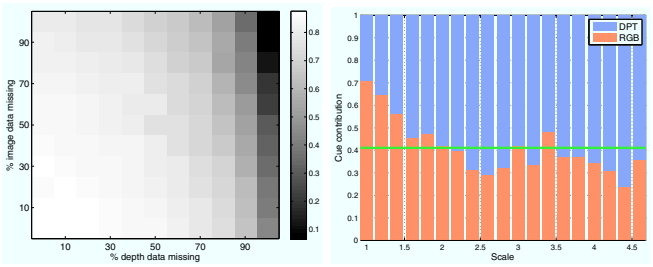

Fig. 4. **Left:** Graceful degradation of classification performance over ranges of invalid inputs. The diagram shows the F-score metric of the HGE fusion scheme with respect to missing data in each modality. Low F-score values are only caused by large amounts of missing information in both modalities. **Right:** Contribution of each modality to the final fused result. Blue bars represent the depth-based people detector contributions, red bars represent the vision-based detector contributions, with the green line being the average contribution over all scales. The visual detector has more importance at smaller scales when depth data are less reliable due to low signal-strength returns or out-of-range readings at far ranges.

source and target feature differences, constitutes a more difficult classification problem.

C. Results on Adaptive RGB-D Fusion

We evaluated our adaptive RGB-D fusion scheme by comparing the results with established late-fusion approaches. For the experiments, we make use of a HOD depth-detector, trained by using the D-CDA procedure shown above, and a HOG image-based detector, trained by using the TUD-Brussels data set [31]. All the quantitative results in this section are computed by using the RGB-D data set. We compute performance results of the following alternative fusion schemes: maximum detectors response (MAX), $\mu_f = \max(h_D(\mathbf{x}), h_R(\mathbf{x}))$; product of the detector responses (AND), $\mu_f = h_D(\mathbf{x}) \cdot h_R(\mathbf{x})$; weighted sum of detector responses (SUM), $\mu_f = \beta_1 \cdot h_D(\mathbf{x}) + \beta_2 \cdot h_R(\mathbf{x})$. All these fusion schemes do not rely on adaptive fusion methods but they make use of fixed rules. The comparative precision-recall curves are shown in Fig. 3-right. The proposed adaptive fusion approach (HGE) is largely the most accurate method, reaching 87.4% EER. The SUM method obtains only 67.8% EER and the MAX method 79.0% EER. The AND method achieves a very low false negative rate at a low detection rate, it fails to cross the precision-recall diagonal, because it returns a detection if and only if both detectors have positive scores. The SUM and MAX methods suffer from the inclusion in the final result of false positive detections from both modalities. All these methods do not balance the relative importance of the modalities with the current data conditions, thus they generally introduce errors when modalities are combined in complex environment situations such as low RGB image contrast or low-return-signal-strength materials.

We then evaluate the contribution of each term of Eq. 4 in the final fusion result. We consider the effect of the gating function $g(\cdot)$, by comparing $g(\cdot)\,h(\cdot)$ with $h_i(\cdot)$, see Fig 3-left. EER values of the classifiers which output has been shaped by $g(\cdot)$ are significantly higher than the unmodified output $h(\cdot)$: EER is 82.1% with respect to 69.9% in the case of depth-based detection, and 77% with respect to 66.8% in the case of image-based detection. We then inspect the effect of $g'(\cdot)$ on Eq. (4) by evaluating the entire fusion architecture (HGE) at different data conditions: we simulated increased missing data in each modality and evaluated the F-score index, see Fig. 4-left. The plot shows a graceful degradation of the HGE fusion scheme that yields low F-score values only when large amounts of missing information occur in both modalities.

It is also interesting to analyze another aspect of the presented fusion scheme. Fig. 4-right shows the average score contribution of each modality evaluated at different scales. Intuitively, the visual detector has more importance at smaller scales when depth data are less reliable due to low signal-strength returns or out-of-range readings at far ranges. The visual detector contribution diminishes at higher scales because statistically a large quantity of valid readings are available at close ranges.

We show qualitative results of the fusion approach on a set collected with a Kinect sensor in a busy train station (Fig. 5). Missing data and invalid readings are present in depth data, motion blur and low contrast in RGB images. The sensor fusion scheme achieves high detection accuracy at a low false negative rate. A HOG image-only detector, using the same images, yields 20% higher misclassifications than the fusion architecture. We also exemplify the capability of fusion with severe local disruption of depth or image data, and show that people are successfully detected in all the cases.

VI. Conclusions

This paper addressed the problem of how to best fuse image and range data for the task of object detection. We proposed two methods that each address issues related to this fusion task: first, an architecture using a hierarchy of mixtures of experts based on Gaussian Processes, able to adaptively weight the output of individual object detectors as a function of missing information and cross-cue data con-

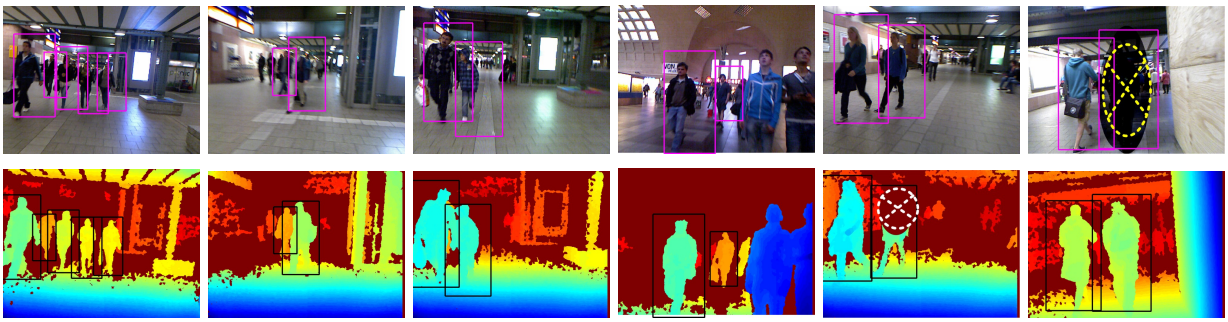

Fig. 5. Qualitative results of the proposed fusion scheme in RGB-D data from a Kinect sensor. The images are collected in a busy train station. Missing data and invalid readings are present in depth data, motion blur and low contrast in RGB images. The sensor fusion scheme achieves high detection accuracy at a low false negative rate. The last two columns contain frames with manually removed portions of the data in both cues (dashed circles). Thanks to the adaptation capability of the system, people are successfully detected in both cases.

sistency. The result is an adaptive strategy that learns when and where to trust which sensory cue. In our experiments using people as object category, we achieved a high detection rate of 87.4% EER, outperforming several established sensor fusion techniques.

The second method, cross-cue domain adaptation, can use annotated image data to improve a depth-based object detector. The method addresses the problem with multi-modal detectors that only few large-scale object data sets exist for RGB-D or 3D range data. In our experiments, we achieved an improvement by more than 13% EER.

Acknowledgements

The authors would like to thank Matthias Luber for the additional data collection. This work has been supported by the German Research Foundation (DFG) under contract number SFB/TR-8.

References

[1] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Eur. Conf. on Comp. Vis. (ECCV)*, 2006.

[2] A. Bergamo and L. Torresani, "Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach," in *Conf. on Neural Inf. Proc. Sys. (NIPS)*, 2010.

[3] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Tran. on Inf. Theory*, vol. 13, pp. 21– 27, 1967.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, 2005.

[5] K. Dietmayer, J. Sparbert, and D. Streller, "Model based object classification and object tracking in traffic scenes from range images," in *IEEE Intelligent Vehicle Symposium*, 2001.

[6] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, 2009.

[7] B. Douillard, D. Fox, and F. Ramos, "Laser and vision based outdoor object mapping," in *Robotics: Science and Systems (RSS)*, 2008.

[8] M. Enzweiler and D. Gavrila, "A multi-level mixture-of-experts framework for pedestrian classification," *IEEE Tran. on Image Proc.*, 2011.

[9] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, 2010.

[10] M. Enzweiler and D. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. on Pattern Analysis & Machine Intell.*, vol. 31, no. 12, pp. 2179–2195, 2009.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge Results (VOC2011)," 2011.

[12] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, 2008.

[13] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, pp. 79–87, 1991.

[14] M. I. Jordan, "Hierarchical mixtures of experts and the em algorithm," *Neural Comput.*, vol. 6, pp. 181–214, 1994.

[15] K. Lai and D. Fox, "Object recognition in 3d point clouds using web data and domain adaptation," *Int. Journ. of Rob. Research*, vol. 29, pp. 1019–1037, July 2010.

[16] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, 2005.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journ. of Comp. Vis.*, vol. 60, no. 2, p. 91, 2004.

[18] M. Luber, K. O. Arras, C. Plagemann, and W. Burgard, "Classifying dynamic objects: An unsupervised learning approach," *Autonomous Robots*, vol. 26, no. 2-3, pp. 141–151, 2009.

[19] *Xbox 360 Kinect Sensor Manual*, Microsoft, Oct 2010.

[20] J. C. Platt, "Probabilities for SV Machines," *Advances in Large-Margin Classifiers*, pp. 61–74, 2000.

[21] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[22] B. Sapp, A. Saxena, and A. Y. Ng, "A fast data collection and augmentation procedure for object recognition," in *AAAI Conf. on Artif. Intell. (AAAI)*, 2008.

[23] G. Schweikert, C. Widmer, B. Schölkopf, and G. Rätsch, "An empirical analysis of domain adaptation algorithms for genomic sequence analysis," in *Conf. on Neural Inf. Proc. Sys. (NIPS)*, 2008.

[24] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, 2011.

[25] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," in *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, 2011.

[26] L. Spinello and K. O. Arras, "People detection in RGB-D data." in *Int. Conf. on Intel. Rob. and Sys. (IROS)*, 2011.

[27] L. Spinello, R. Triebel, and R. Siegwart, "Multiclass multimodal detection and tracking in urban environments," *Int. Journ. of Rob. Research*, vol. 29, no. 12, pp. 1498–1515, 2010.

[28] L. Spinello, K. O. Arras, R. Triebel, and R. Siegwart, "A layered approach to people detection in 3D range data," in *AAAI Conf. on Artif. Intell. (AAAI)*, 2010.

[29] A. Teichman, J. Levinson, and S. Thrun, "Towards 3d object recognition via classification of arbitrary object tracks," in *Int. Conf. on Rob. & Autom. (ICRA)*, 2011.

[30] R. Triebel, J. Shin, and R. Siegwart, "Segmentation and unsupervised part-based discovery of repetitive objects," in *Robotics: Science and Systems (RSS)*, 2010.

[31] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, 2009.

[32] J. Wu, C. Geyer, and J. M. Rehg, "Real-time human detection using contour cues," in *Int. Conf. on Rob. & Autom. (ICRA)*, 2011.