2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Unconstrained 1D Range and 2D Image Based Human Detection

Mehmet Kocamaz[1] and Fatih Porikli[2]

Abstract— **An accurate and computationally very fast multi-modal human detector is presented. This 1D+2D detector fuses 1D range scan and 2D image information via an effective geometric descriptor and a silhouette based visual representation within a radial basis function kernel support vector machine learning framework. Unlike the existing approaches, the proposed 1D+2D detector does not make any restrictive assumptions on the range scan positions, thus it is applicable to a wide range of real-life detection tasks. To analyze the discriminative power of the geometric descriptor, a range scan only version, 1D+, is also evaluated. Extensive experiments demonstrate that the 1D+2D detector works robustly under challenging imaging conditions and achieves several orders of magnitude performance improvement while reducing the computational load drastically.**

In addition, a new multi-modal (LIDAR, depth image, optical image) dataset, *DontHitMe*, is introduced. This dataset contains 40,000 registered frames and 3,600 manually annotated human objects. It depicts challenging illumination conditions in indoors and outdoors environments and is publicly available to our community.

I. INTRODUCTION

According to National Highway Traffic Safety Administration reports [1], thousands of pedestrians lose their lives in traffic accidents. At least the same number of pedestrians get injured and become handicapped because of these incidents. Incorporating advanced human detection solutions in intelligent driving systems would significantly reduce these unfortunate events.

There are two sets of challenges that make the human detection task complicated. The first one is the external factors. These factors are not object depended and often caused by environmental elements. Illumination variations, insufficient street lighting, saturation due to headlights, cast shadows, reflections, weather conditions, existence of human-like objects and clutter, and imaging noise fit into this category. External factors have absolute effects to the performance of the detection process.

The second set of challenges are due to the human itself, thus may be called as the internal factors. Humans have articulated body parts that move, rotate, and deform. They stand up, walk, run, bend and make body gestures. The appearance, height, weight, and clothing might differ significantly from one to another. Therefore, their bodies appear in different shapes and silhouettes. In addition, human body has various poses from distinct view points. All these factors make the objective of the human detection considerably more difficult than detection of rigid objects.

[1] University of Delaware
[2] Mitsubishi Electric Research Labs

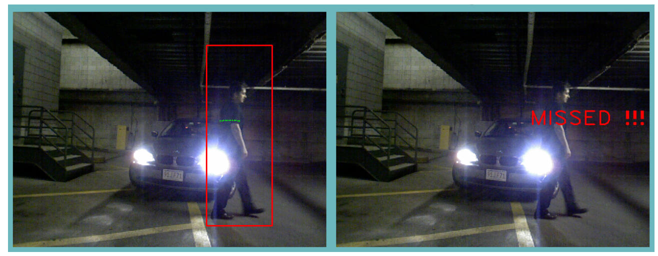

Fig. 1. Left image shows the detected pedestrian by the proposed multi-modal human classifier under severe illumination conditions. Single-modal classifiers [2] (and conventional multi-modal approaches) are not be able to detect as can be seen in the right image.

This paper presents a novel multi-modal human detector that fuses 1D range scans from a LIDAR (Laser Imaging Detection And Ranging) sensor and 2D monocular images from an optical camera. The proposed algorithm integrates the photometric and depth features obtained from both data modalities in a joint classifier. It is robust under difficult environmental conditions. Unlike the existing approaches, it can detect humans even if the range scan beams hit upper torso and head of the body without making any assumptions about the visibility of the legs. This is critical for real world applications. For instance, the scan beam may easily miss the legs when the road climbs over a hill or there is a short subject, e.g. a child, in the detection range. The legs can be occluded due to skirts, bags, strollers, etc. When the pedestrian stands up sideways, only one leg is visible. Fusing multiple modalities not only increases the detection accuracy for such examples but also improves the computational time. Since it efficiently narrows down the search region in the image, our detector runs very fast.

This work makes several improvements to the human detection problem in the following ways:

1: A highly accurate and computationally fast multi-modal human detector that fuses 1D range scans and 2D images is presented. This 1D+2D detector does not make any restrictive assumptions about the range scan positions.

2: A simple yet effective geometric descriptor is introduced for LIDAR data. A single-modal human detector, 1D+, using this descriptor is developed. This detector achieves higher accuracy than the state-of-the-art human classifiers based on 1D range scans.

3: It is shown that the multi-modal classifier can be trained with less precise range information, for instance using Kinect sensor depth data, to eliminate the need for expensive and cumbersome manual labeling.

4: A new LIDAR, camera, and Kinect sensor based, regis-

978-1-4673-6357-0/13/$31.00

tered human dataset, called as *DontHitMe*, is made available. This dataset consists of 40,000 frames, 450 people. and 3,600 ground positive samples (about one per ten frames). It is one of the largest publicly available multi-modal human datasets.

A review of the related work is summarized in the next section. Descriptors, fusion modules, training algorithm and the proposed classifiers are described in Section 3. The details of the human dataset collection process, and the ground truth generation are explained in Section 4. In the same section, experimental results of both single- and multi-modal classifiers are analyzed. Finally, the future directions of this work are drawn in the last section.

II. RELATED WORK

Computer vision and robotics communities have been conducting extensive research on human detection for years. In both fields, selected sensor types has a direct impact on the fundamentals of the developed method. Basically, human detection algorithms can be categorized in two groups based on the modality of the input data as explained below.

A. Single Modal Human Detection

There are two essential sensor types used for single modal detection. First group includes the visual sensors, such as monocular cameras. Sensors that provide 3D geometric cues, such as one or multi layer LIDAR and Kinect, form the second group.

Visual human detectors take an input image, compute descriptors within all possible subwindows and ask a classifier to determine whether there is a human inside the subwindows or not. In earlier image based human detection works [3], [4], Haar wavelets are used to construct descriptors and train multiple linear Support Vector Machines (SVMs).

A seminal human detection technique that uses the Histogram of Oriented Gradients (HOG) features is proposed in [2]. For speed improvement, a rejection cascade of AdaBoost classifiers using the HOG features is described in [5]. The region covariance features (COV) are first introduced in [6] and a classifier based on the underlying Riemannian manifold is deployed in [7]. These holistic methods achieve remarkable results, but they may suffer from occlusions.

Alternatively, human detection can be done by identifying body parts and their common shapes [8], [9], [10], [11], [12]. In these methods, local features for body parts are determined and combined to form human models. In [13] and [14] human silhouette information is also taken into account. These methods are more robust to occlusion. However, their performance highly depends on the image resolution of the human body parts.

Detectors that rely only on geometric cues often extract features from 3D or range scan data. For example, [15] applies a set of oriented filters to the spatial depth histograms. Instead of a classifier, a simple threshold operation is performed to find the humans. Depth images are converted to 3D point clouds in [16], [17]. A dictionary is constructed from the geodesic local interest points by [16]. This method has a high detection rate as long as humans are not occluded and touch other objects. A large feature vector that employs the histograms of the local depth information is used to represents humans [17]. This approach is robust to occlusions yet it is computationally very demanding and not suitable for real time applications. Only a single LIDAR range scan is processed to form a leg descriptor in [18] and [19]. These approaches extract a number of predefined features from the segmented line parts and train classifiers. They can detect humans if the legs are visible and the LIDAR beam hits at the lower torso level.

Integrated human detection and tracking solutions that use 3D data from Velodyne LIDAR are described in [20], [21]. These methods are more accurate than [18] and [19], yet the comparably expensive cost of the sensor limits their applicability.

B. Multi-Modal Human Detection

The underlying idea of using multiple modalities is to combine their complementary advantages.

Multi-modal detection algorithms can be centralized and cascaded. Centralized approaches combine the features obtained from different sensors in a single feature vector [22] and train a single classifier.

Cascaded approaches construct multiple descriptors and train separate classifiers for each modality. They compute classifier confidences [23], [24] or impose one of the classifier to reduce the search space of the other classifier [25]. The classifiers explained in [22], [23], [24] use 1D range scans and color images to construct features and extract features. [23] and [22] use the HOG and COV features, whereas [24] uses Haar-like features to form their visual descriptors. Similar to [18], they cannot handle the situations where the range scans hit human body other than the legs.

The method described in [26] utilizes 3D features obtained from image and 3D point cloud. However, it is computationally expensive to retrieve the 3D geometric features for realtime applications. [25] shares a similar concept and focuses on reducing the computational load. This method uses 3D information retrieved from stereo images to limit the search. Since it has no geometric feature extraction or information fusion mechanism, it still suffers from abrupt changes in the illumination conditions.

III. 1D+2D DETECTOR

To take the advantages of the geometric and visual information, our 1D+2D multi-modal human detector combines the range scan and image descriptors into a single representation. It works in the joint higher-dimensional feature space. A diagram of the classifier is given in Fig. 2.

For a training image window W_i, the corresponding 1D range scan segment $L_i = (d_1, ..., d_{m_i})$ within the window can be obtained either from the LIDAR or from the depth camera. In the case of the LIDAR sensor, there is a single, horizontal, synchronously acquired range scan segment within the window. On the other hand, the depth camera can provide multiple horizontal range scan segments, which are particularly valuable for training. Here, d is the depth, i.e.

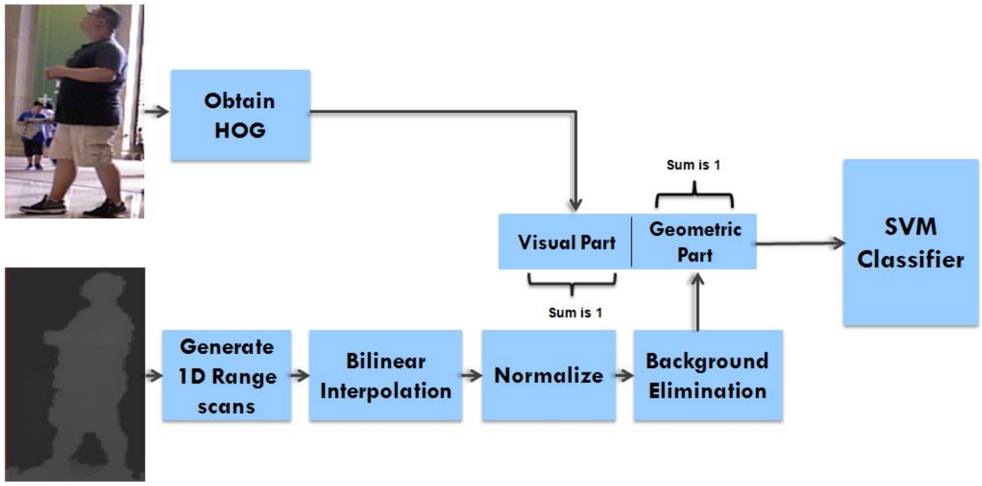

Fig. 2. Training process of the 1D+2D detector.

the distance of the sensor to a scene point in camera normal direction.

A. Geometric Descriptor

In contrast to [18] that assumes the geometric descriptor corresponds to leg region, our geometric descriptor f^{1D} applies to every part of the human body. It is obtained by the following procedure:

1) Depending on the size and depth of the human objects, range scans L_i for positive samples form arbitrary length vectors

$$f_i^{1D} = [d_1, ..., d_{m_i}]_i^T, \quad 1 \leq m_i \leq \max(\|w_W\|) \quad (1)$$

where $\|w_W\|$ is the width of the window. In order to map the arbitrary length feature vectors onto a uniform, fixed dimensional feature space R^m, an m-point bilinear interpolation, B_m, is performed on f_i^{1D}. After the interpolation, the dimension of f_i^{1D}, that is m_i, becomes m

$$f_i^{1D} = [d_1, ..., d_m]_i^T \leftarrow B_m(f_i^{1D}). \quad (2)$$

2) The distance between the sensor setup and a human differs significantly in the scene. To compensate for this distance, the closest point depth, d_C, in f_i^{1D} to the sensor setup is determined. Then, d_C is subtracted from f_i^{1D}.

$$d_C = min(d_1, ..., d_m), \quad d_C \neq 0 \quad (3)$$

$$f^{1D} \leftarrow f^{1D} - d_C = [d_1 - d_C, ..., d_m - d_C]^T. \quad (4)$$

3) Human objects stand in front of all kinds of backgrounds. Background clutter, as well as other objects in the scene, may be positioned at different distances from the human objects. This causes considerable geometric feature variation around the silhouette of the human body. Capturing all this variation in the training data would be one approach. Yet, this requires a huge amount of training data, which would be impractical. Besides, it may cause the classifier to fail because of the weakened discriminative power of the descriptors.

Therefore, the depth values of the feature vector elements that are above a human shape threshold are upper bracketed. The threshold, d_H, is set to the maximum possible radius of a human. If a point in the feature vector f^{1D} has a depth value larger than the threshold, it is set to the maximum radius. As a result, the variation due to the other objects and background clutter are eliminated effectively:

$$d_k = \begin{cases} d_H & \text{if } d_k \geq d_H \\ d_k & \text{otherwise} \end{cases} . \quad (5)$$

B. Visual Descriptor

Due to its shape representation ability, computational simplicity, and robustness to illumination changes up to a certain degree, the HOG feature is used to form the visual part of our human descriptor, $f^{2D} = [v_1, ..., v_n]^T$ in the classifier. The HOG features can represent efficiently the local appearance by a distribution of the edge gradients in a cell within an image region. These cells, either overlapping or on a regular grid, are smaller components of an image window. A histogram is obtained within a cell. These local cell histograms are concatenated into a larger window descriptor. All cell histograms of the window descriptor are normalized using the accumulated energy within the window for additional illumination robustness.

C. Combined Descriptor & Classifier Training

The geometric f^{1D} and visual f^{2D} features are concatenated in the same feature vector to form the final multi-modal human descriptor, f.

The raw geometric and visual feature vectors have different dimensions, thus their individual contributions in the combined multi-modal descriptor are not balanced. To overcome this issue, individual vectors are normalized to unit norm:

$$f^{1D} \leftarrow \frac{f^{1D}}{\sum_{k=1}^{m} d_k} \quad (6)$$

and

$$f^{2D} \leftarrow \frac{f^{2D}}{\sum_{k=1}^{n} v_k}. \quad (7)$$

The combined descriptor in R^{m+n} is then $f = [f^{1D} f^{2D}]^T$.

In training, the negative samples are chosen from the windows where there are no human objects. Since the window size changes according to the depth value of the window center, the size variation of the negative samples comes naturally. Even though in practice only LIDAR sensor data is available with the image, our training process still benefits from the additional depth camera data.

We use Support Vector Machines (SVMs) as our base classifiers. SVMs fits a hyperplane between the positive and negative training samples in the feature space. The decision boundary is defined by a set of support vectors that separate the positive and negative samples in a maximum margin. The decision function of SVM is

$$h(f) = \sum_{i=1}^{m} \alpha_i [\phi(f).\phi(f_i^*)] \tag{8}$$

where α_i are the weight of the corresponding m support vectors f_i^* and ϕ is a mapping function to a space $\mathcal{H}$. The dot products in the decision function can be replaced by a kernel function:

$$k(f, f_i^*) = \phi(f).\phi(f_i^*) \tag{9}$$

By using a kernel function the classifier becomes a hyperplane in $\mathcal{H}$, yet it may be non-linear in the original input space. For given a set of labeled samples (x_i, y_i) where the labels $y_i = \{-1, 1\}$, the learning problem of SVM can be formulates as the minimization of

$$\min_{w,\varepsilon,b} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \varepsilon_i \tag{10}$$

subject to

$$y_i(w.f_i - b) \geq 1 - \varepsilon_i, \quad \varepsilon_i \geq 0 \tag{11}$$

where ε_i a penalty for the misclassified samples. The above optimization tries to classify as many training sample as possible correctly. Also, the minimization of $\|w\|$ makes the margin as large as possible. C is a variable term to set the relative influence.

We use the Radial Basis Function (RBF) as the kernel function of SVM:

$$\phi(f).\phi(f_i^*) = exp(-\gamma\|f - f_i^*\|^2) \tag{12}$$

where γ is the width of Gaussian kernel width. By using RBF, it is always possible to find a decision function that perfectly represents a shape in a higher, possibly infinite, dimensional space. By incorporating RBF, SVM decision function takes the final form of

$$h(f) = \sum_{i=1}^{m} \alpha_i exp(-\gamma\|f - f_i^*\|^2) \tag{13}$$

the result of final classification is the sign of $h(f)$. This decision function depends on the distance between the support vectors and the data, thus normalizing the geometric f^{1D} and visual f^{2D} feature vectors to unit norm, as formulated in Eqns. 6 and 7, is necessary. Otherwise, higher dimensional features would be favored by the SVM decision function.

In addition to the above 1D+2D detector, a single-modal classifier, called as 1D+ detector, is also trained by SVM using only the 1D range scans to assess the discriminative power of the proposed geometric descriptor.

D. Fast Detection

Since the speed of the human detection is an important factor, the 1D+2D detector is employed in a joint fashion that takes advantage of the depth information to eliminate the unnecessary window evaluations.

To determine whether a test window depicts a human, the corresponding 1D and 2D features are computed on the registered data. The range scan line L is aligned with the 2D image I by a perspective transformation $L_I : T(L)$ to obtain a set of image pixel coordinates $L_I = (p_1, ..., p_n)$ in the image.

A search window $W(x, y, \delta x, \delta y)$ centered around p_k is slided on the coordinates of L_I. The size (width δx and height δy) of W is set according to the original depth value d_k of the point p_k such that for smaller depth values (objects closer to the sensor setup) the window size becomes larger. The window size is also proportional to the average human size at the corresponding original depth value.

There is no guarantee that the LIDAR beam always hits a specific level of the human body in a real application, thus the vertical position y of the image window W is not fixed. Instead, multiple windows at different vertical positions $y \pm \Delta y_j$ are tested for each p_k. Similar to the selection of the window size, the number of the vertical windows and their separation are determined by the original depth value of the center point. In this case, if d_k has a large value, a smaller vertical jumps Δy_j between multiple windows is desirable.

Within each window, the geometric descriptor f^{1D} and visual descriptors f^{2D} are computed, normalized, and concatenated into f. If the sign of the $h(f)$ in the SVM classifier is positive, a human is detected by the multi-modal classifier. Algorithm 1 outlines the testing procedure.

In contrast to the conventional visual-only human detectors that need to search entire image at different scales, our 1D+2D classifier reduces drastically the search space. It eliminates completely the image scaling step. Using L_I help to prune most of the image areas, which decreases the computational load greatly.

In practice, window evaluations can be ordered from nearest to far based on the LIDAR sensor depth values to determine the most critical object first.

IV. Dataset and Experiments

A. DontHitMe Dataset & Sensor Setup

In supervised learning, the quality and quantity of training data are very critical for the final performance of the classifier. More training data prevents from overfitting, improves generality, and enables trained models to capture possible variations of target class samples. Since our purpose is to construct an inclusive and unconstrained classifier that

Algorithm 1 Detection Algorithm

Inputs: $L = (d_1, ..., d_n)$ range scan points, I, T, h

* Compute L_I, by $L_I : T(L)$
for k=1, ... , n (all points in L_I)
* Scale search window W by $1/d_k$
* Compute geometric descriptor $f^{1D} = [d_1, ..., d_m]^T$ inside W using Eqs. 1-5
* Determine, Δy_j, vertical jump offsets from d_k
for each Δy_j for W
  * Compute HOG $f^{2D} = [v_1, ..., v_n]^T$
  * Normalize f^{1D} and f^{2D}
  * Concatenate f^{1D} and f^{2D} to $f = [f^{1D} f^{2D}]^T$
  * Compute $h(f) = \sum_{i=1}^{m} \alpha_i exp(-\gamma \|f - f_i^*\|^2)$
  * if $h(f) > 0$ detect human, remove underlying points from L_I

performs accurately without making any assumption about the range scan position on the human body, a large number of training samples is required for training. However, it is cumbersome to collect such a large number of registered LIDAR and camera data where range scans hit humans on different parts of their bodies. To capture different pose, appearance variations and scan positions, the height and position of the LIDAR must be modified excessively. This is definitely a tedious and inefficient task with no guarantee of capturing sufficient amount and quality of data.

To our advantage, it is possible to generate a high number of diverse range scans for positive and negative samples by using a depth camera that provides the 3D structure of the scene. Any number of scans can be obtained from a depth image by converting the geometric information into LIDAR-like readings synthetically.

Towards this goal, a sensor setup composed of an Asus Xtion Pro Live IR and color camera, and a Hokuyo URG-04LX LIDAR was used. Three sensors, IR camera, color camera and LIDAR were registered in the same coordinate system. A multi-modal human data set, called as *DontHitMe*, was collected in outdoors (parking lots, streets, etc.) and indoors (campus, etc.) buildings. Since the IR camera is sensitive to the sunlight, outdoor data was recorded when there was no direct sunlight in the scene. In addition to the color and depth images, this dataset also includes registered 1D LIDAR range scans. It contains 40,000 images of 450 different humans in different poses, appearance variations, lighting conditions, and shadow artifacts. Several human shapes that present a challenge to existing human classifiers, such as women in skirts and small children were recorded. To capture the variance of the human poses, images are recorded sequentially at 8 fps. The location and height of the setup was changed during the collection process to collect samples in different backgrounds. Modifying the height of the sensor setup was diversified the recorded 1D range scans.

The original LIDAR range scans hit human body on different parts from the legs to the head. A total of 3,600 manual ground truth positions in images, depth camera data, and range scans were annotated. Each human in the dataset was labeled with a bounding box, $W(x, y, \delta x, \delta y)$. *DontHitMe* dataset is divided into two different categories. The first dataset, called as *DontHitMe-Indoor*, includes 30,000 frames and 3,000 ground truths which are recorded indoor campus buildings. The second dataset is collected in outdoors at night times and contains more challenging cases for human detectors, such as insufficient lighting and severe illumination changes because of car headlights. This dataset contains 10,000 frames and 600 ground truths, called as *DontHitMe-Night*.

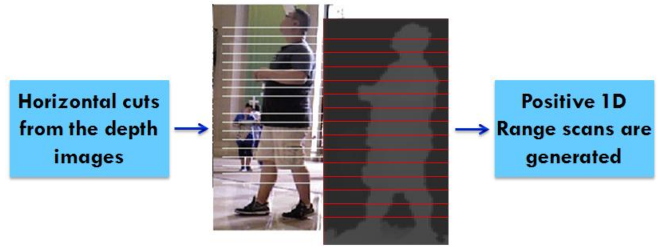

Fig. 3. 1D range scans are generated from the depth camera data for each positive window.

To complement the original LIDAR data, the depth camera data in *DontHitMe-Indoor* were processed to obtain additional synthetic range scans as shown in Fig. 3. These horizontal scans were produced by uniformly sampling multiple positions vertically along the labeled human window W for the positive samples. In this way, multiple scans were generated from each part of human body, from the legs to the head. A depth scan $L_i = (d_1, ..., d_{m_i})_i$ was discarded if it contained points where the depth camera does not provide a valid distance.

B. Experiments

Several experiments were conducted to quantify the performance of the proposed multi-modal human classifier, 1D+2D, and its range scan only version, 1D+.

In the first experiment, we analyzed the performance of 1D+ detector. We obtained 46,000 positive and 376,000 negative samples from the LIDAR sensor scans and depth images of *DontHitMe-Indoor* dataset. A total of 43,000 positive samples are generated synthetically from the depth images by uniform sampling and additional 3,000 positive samples were obtained from the recorded 1D range scans.

In order to reduce the variability in the testing scores, we performed multiple rounds of 10-fold cross-validation. We aimed to see the performance of the 1D+ detector at the different parts of the human body. Therefore, the positive samples in *DontHitMe-Indoor* dataset are divided into 3 categories, as upper body, torso and lower body.

The outcomes of the proposed and the existing state-of-the-art classifiers for the separate human body parts and for negative samples can be seen in Table I. The results are compared to [18], which is a 1D range scan based human classifier. As visible, our 1D+ detector outperforms [18] at least by 20.2% for each part of the human body. The result

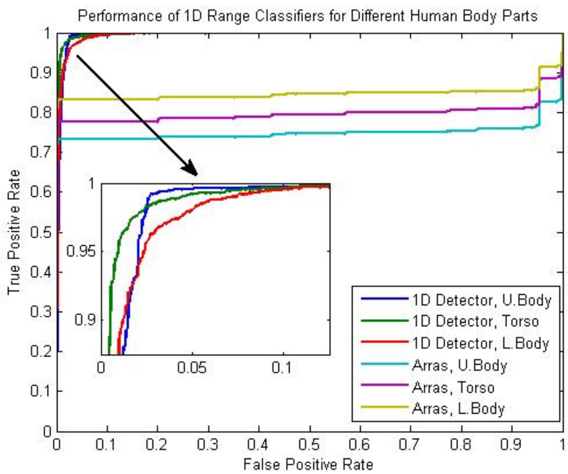

Fig. 4. ROC curves of the 1D+ Detector and Arras's classifier [18] at different parts of the human body.

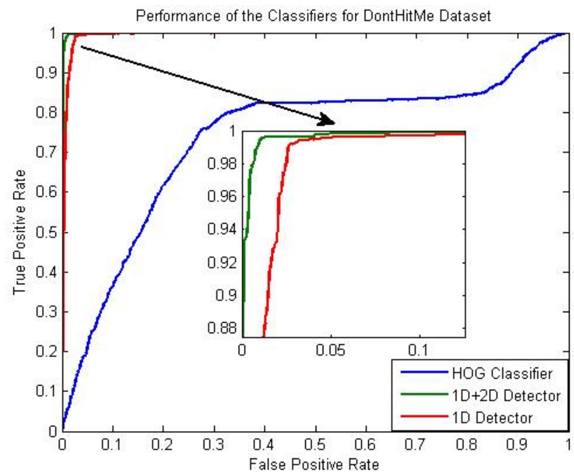

Fig. 5. Performance of the benchmark HOG [2] and the proposed 1D+2D and 1D+ human classifiers tested on *DontHitMe* dataset.

TABLE I

COMPARISONS OF 1D RANGE SCAN BASED HUMAN DETECTORS FOR DIFFERENT HUMAN BODY PARTS

Test Set	1D+ Detector	Arras et al. [18]
Upper Body	97.5%	78.6%
Torso	97.9%	82.7%
Lower Body	96.8%	86.6%
Negative Samples	96.5%	5.6%

of this experiment shows that assumptions on the visibility of the legs is not valid for real-life scenarios. The 1D+ is more robust and achieves remarkable accuracy at each level of the human body as can be seen in the detection performance curves of the classifiers in Fig. 4. As expected, the method explained in [18] shows its best performance if the range scans hit the lower part of the human body. Whereas the performance of our detector is almost same at different parts of the body. Proposed 1D+ does not miss any human at 89% false detection level. One of the main reasons of the consistent performance of our classifier at each part is that the positive samples are provided to our detector uniformly from different body parts in the training phase. Also, it learns more diverse geometric cues from every different part of the the body from head to the feet.

Another experiment was conducted to measure the performance of the proposed 1D+2D detector. A total of 1,000 positive and 10,000 negative visual descriptors were obtained from *DontHitMe-Indoor* dataset. For each visual descriptor, 20 different geometric descriptors were generated synthetically from different parts of the body by uniformly sampling in their corresponding depth images. In this way, total of 20,000 positive and 200,000 negative multi-modal samples which merge visual and geometric descriptors were generated from *DontHitMe-Indoor* dataset to train the 1D+2D detector. Also, for comparison purposes, 1D+ detector was trained only with the geometric descriptors and the HOG human classifier [2] was trained with the visual descriptors of this set. The accuracy of the proposed 1D+2D detector and 1D+ detector were compared to the HOG human classifier. As in the previous test, multiple 10-fold cross-validations were performed. During this experiment, it was ensured that the test fold and training folds include the samples obtained from different humans. In this way, testing of the geometric and visual descriptors obtained from the same positive samples used in training are prevented. The ROC curves of this experiment can be seen in Fig. 5. The 1D+2D detector and 1D+ detector perform significantly better than the visual only detector.

The proposed classifiers were tested with 600 labeled ground truth images of *DontHitMe-Night* dataset to quantify the performance of the classifiers under severe illumination conditions in outdoor. In this experiment, the classifiers trained in the previous experiment were applied on the night dataset. No new classifier was trained by using *DontHitMe-Night* and no syntectic range scans were generated from the depth images of this dataset. The tested geometric human descriptors were obtained only from the recorded LIDAR scans. The ROC curves of the 1D+2D, 1D+, and [2] detectors are displayed in Fig. 6. It can be seen that the HOG descriptor is not enough to represent the human under insufficient lighting and at night times. Our single-modal human descriptor achieved better accuracy than the HOG descriptor. Fusing the visual and geometric cues in a joint feature vector helped to improve the performance; 1D+2D detector outperforms consistently the other alternatives.

Since our geometric descriptor is obtained from the LIDAR scans, our 1D+2D detector is more capable of handling image motion blur than the HOG classifier. Such motion blur examples can be seen in Fig. 5, for example, at the foot level of the pedestrians.

Note that, since it is accurate and computationally feasible at the same time, we compare against the HOG detector that uses SVM-RBF [2] for the most objective evaluations. There are other visual features that can generate higher

TABLE II
AVERAGE RUNNING TIME AND FALSE ALARM RATE (AT 95% TRUE DETECTION RATE) OF DIFFERENT CLASSIFIERS FOR *DontHitMe-Night* DATASET.

Classifier	Time (in sec)	FAR at 0.95 TDR
HOG [2]	0.6	86%
1D+ Detector	0.0002	0.5%
1D+2D Detector	0.05	0%

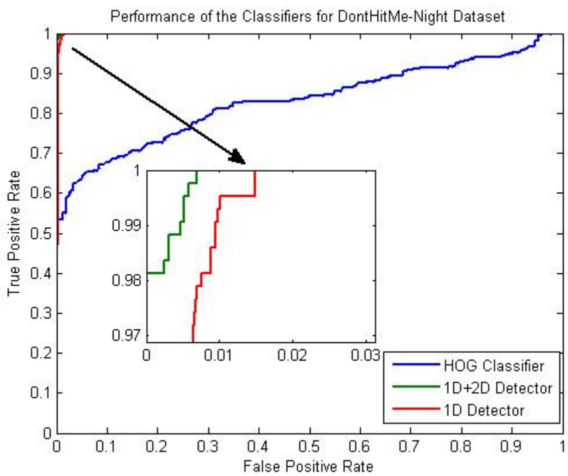

Fig. 6. ROC curves of the classifiers for *DontHitMe-Night* dataset.

detection results. Yet, such methods have prohibitively high computational loads for most practical applications.

C. Computational Load

A 64x128 detection window size was chosen for both the HOG and the proposed 1D+2D detector in the experiments. The dimension, m, of geometric feature f^{1D} is set to 40. The visual feature, f^{2D}, has the dimension of 3780. We used a machine which has 32GB RAM and Intel i7-2760QM quad processor to train and test the classifiers. The classifiers are implemented in native C++ language of Visual Studio 2010 Pro. The training phase of the 1D+2D detector consumed the largest memory among the classifiers in the second experiment since it requires 220,000 descriptors to fit into 29GB RAM, which took $\sim$5 hours.

We compare the computational time and accuracies of the classifiers for *DontHitMe-Night* dataset experiment as can be seen in Table II. The average processing time of a 640$\times$480 scale-space image (10,000 detection windows) by the benchmark HOG classifier is about 0.6 second. At 95% true detection rate, false alarm rate of it is 86%, whereas the false alarm rate of the 1D+2D detector is 0 on the tested dataset. Since the search space of the 1D+2D detector is reduced efficiently by the factors explained above, its average processing time is just 0.05 second. The proposed geometric descriptor has much less dimensions in comparison to other descriptors and it is easy to compute. Thus, 1D+ detector was be able to run at 0.0002 second per scan in the same experiment.

V. CONCLUSION

We present an accurate and computationally very fast multi-modal human detector. This 1D+2D detector combines 1D range scan and 2D image information within a SVM-RBF framework. Unlike the existing approaches, the proposed 1D+2D detector does not make any restrictive assumptions on the range scan positions, thus this unconstrained detector is applicable to a wide range of real-life detection tasks. We also discuss a range scan only version 1D+.

Our extensive experiments demonstrate that the 1D+2D detector works robustly under challenging imaging conditions and achieves several orders of magnitude performance improvement (99% true detection at 0.005% false alarm rate in comparison to 54% true detection at 0.005% same false alarm rate on the benchmark) while reducing the computational load drastically (from 0.6 sec to 0.05 sec).

In addition, a new multi-modal (LIDAR, depth image, optical image) dataset, *DontHitMe*, is introduced. This dataset contains 40,000 registered frames and 3,600 manually annotated human objects. It depicts challenging illumination conditions in indoors and outdoors environments and will be publicly available to our community.

As future work, the presented approach can be extended to multi-class problems.

REFERENCES

[1] "Traffic Safety Facts Annual Reports, National Center for Statistics and Analysis," http://www-nrd.nhtsa.dot.gov/CATS/listpublications.aspx?Id=E, October 2012.
[2] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2005.
[3] Constantine Papageorgiou and Tomaso Poggio, "A trainable system for object detection," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, Jun 2000.
[4] Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio, "Example based object detection in images by components," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 23, pp. 349–361, 2001.
[5] Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, and Shai Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2006.
[6] Oncel Tuzel, Fatih Porikli, and Peter Meer, "Region covariance: a fast descriptor for detection and classification," in *European Conference on Computer Vision, (ECCV)*, 2006.
[7] Oncel Tuzel, Fatih Porikli, and Peter Meer, "Human detection via classification on Riemannian manifolds," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2007.
[8] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, Jan 2005.
[9] S. Ioffe and D. A. Forsyth, "Probabilistic methods for finding people," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 45–68, Jun 2001.
[10] Remi Ronfard, Cordelia Schmid, and Bill Triggs, "Learning to parse pictures of people," in *European Conference on Computer Vision, (ECCV)*, 2002.
[11] Krystian Mikolajczyk, Bastian Leibe, and Bernt Schiele, "Multiple object class detection with a generative model," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2006.
[12] Krystian Mikolajczyk, Cordelia Schmid, and Andrew Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *European Conference on Computer Vision, (ECCV)*, 2004.

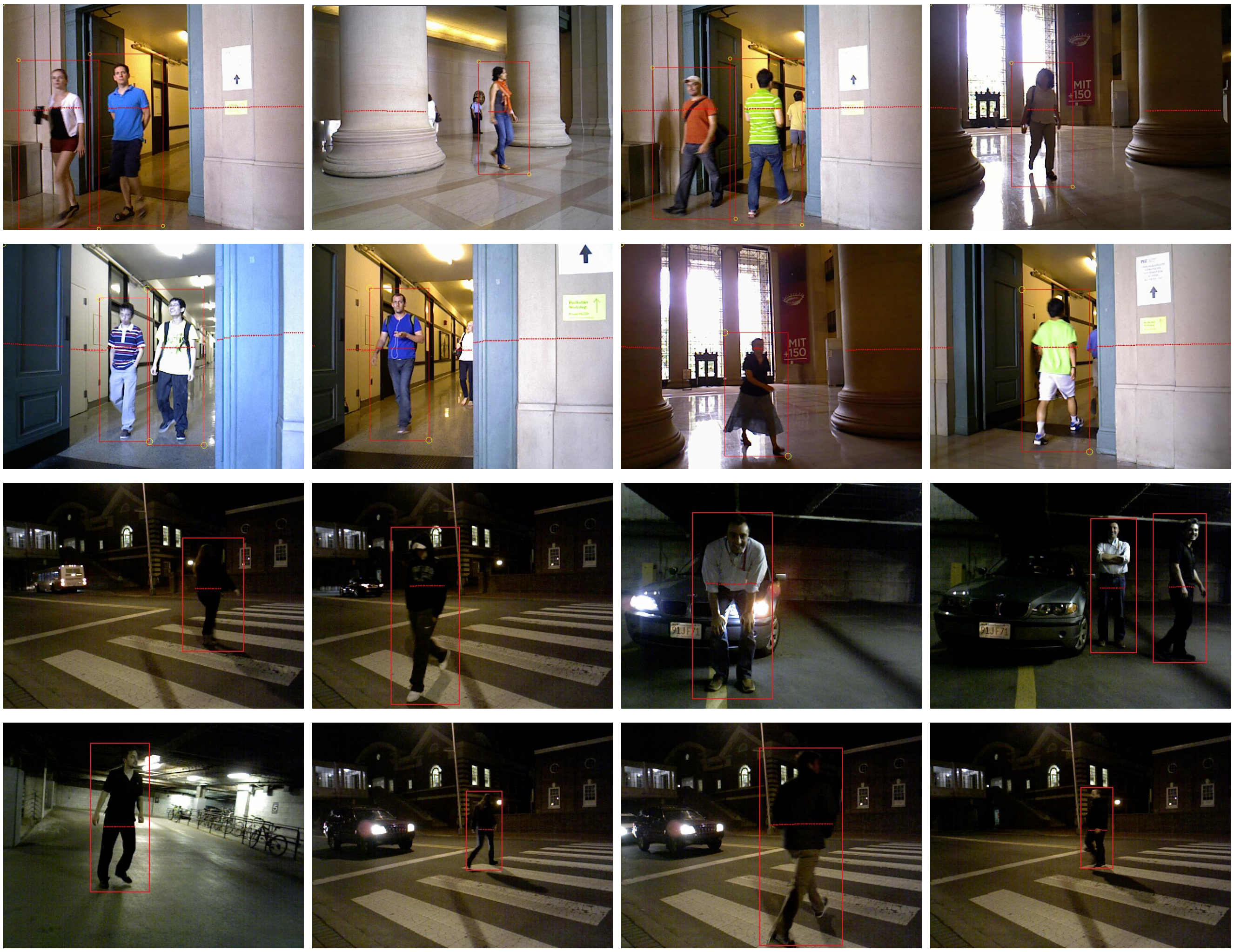

Fig. 7. Sample results of the 1D+2D multi-modal human detector. First two rows display detections in sample images from DontHitMe-Indoor dataset. Last two rows show sample results that were missed by the HOG based SVM-RBF [2] but accurately detected by the 1D+2D in DontHitMe-Night dataset.

[13] Dariu Gavrila and Vasanth Philomin, "Real-time object detection for smart vehicles," in *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, 1999.

[14] Andreas Opelt, Axel Pinz, and Andrew Zisserman, "Incremental learning of object detectors using a visual shape alphabet," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2006.

[15] Taewan Kim, Sangho Cho, Jongmin Yoon, and Daijin Kim, "Pose robust human detection in depth image using four directional 2d elliptical filters," in *IEEE International Symposium on Multimedia, (ISM)*, 2009.

[16] Christian Plagemann, Varun Ganapathi, Daphne Koller, and Sebastian Thrun, "Real-time identification and localization of body parts from depth images," in *IEEE Int. Conf. on Rob. and Autom. (ICRA)*, 2010.

[17] Sho Ikemura and Hironobu Fujiyoshi, "Real-time human detection using relational depth similarity features," in *Asian Conference on Computer vision, (ACCV)*, 2010.

[18] Kai O. Arras, Oscar Martinez Mozos, and Wolfram Burgard, "Using boosted features for the detection of people in 2D range data," in *IEEE Int. Conf. on Rob. and Autom. (ICRA)*, 2007.

[19] Christiano Premebida, Oswaldo Ludwig, and Urbano Nunes, "Exploting LIDAR-based features on pedestrian detection in urban scenarios," in *IEEE Conference on Intelligent Transportation Systems, (ITSC)*, 2009.

[20] Luis Ernesto Navarro-Serment, Christoph Mertz, Nicolas Vandapel, and Martial Hebert, "LADAR-based pedestrian detection and tracking," in *Workshop on Human Detection from Mobile Robot Platforms, (ICRA)*, 2008.

[21] Bharath Kalyan, K. W. Lee, W. Sardha Wijesoma, D. Moratuwage, and Nicholas M. Patrikalakis, "A random finite set based detection and tracking using 3d LIDAR in dynamic environments," in *IEEE International Conference on Systems, Man and Cybernetics, (SMC)*, 2010.

[22] Cristiano Premebida, Oswaldo Ludwig, and Urbano Nunes, "LIDAR and vision-based pedestrian detection system," *Journal of Field Robotics*, vol. 26, no. 9, pp. 696–711, Sep 2009.

[23] Luciano Spinello and Roland Siegwart, "Human detection using multimodal and multidimensional features," in *IEEE Int. Conf. on Rob. and Autom. (ICRA)*, 2008.

[24] Z. Zivkovic and B. Krose, "Part based people detection using 2D range data and images," in *IEEE International Conference on Intelligent Robots and Systems, (IROS)*, 2007.

[25] Rodrigo Benenson, Markus Mathias, Radu Timofte, and Luc J. Van Gool, "Pedestrian detection at 100 frames per second," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2012.

[26] Stephen Gould, Paul Baumstarck, Morgan Quigley, Andrew Y. Ng, and Daphne Koller, "Integrating visual and range data for robotic object detection," in *ECCV workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2)*, 2008.