Semantic fusion of laser and vision in pedestrian detection

Luciano Oliveira *, Urbano Nunes, Paulo Peixoto, Marco Silva, Fernando Moita

Institute of Systems and Robotics, University of Coimbra, Pinhal de Marrocos, Polo II, Coimbra, Portugal

A R T I C L E   I N F O

Article history:
Received 7 December 2009
Received in revised form 31 March 2010
Accepted 7 May 2010

Keywords:
Semantic sensor fusion
Pedestrian detection
Markov logic network

A B S T R A C T

Fusion of laser and vision in object detection has been accomplished by two main approaches: (1) independent integration of sensor-driven features or sensor-driven classifiers, or (2) a region of interest (ROI) is found by laser segmentation and an image classifier is used to name the projected ROI. Here, we propose a novel fusion approach based on semantic information, and embodied on many levels. Sensor fusion is based on spatial relationship of parts-based classifiers, being performed via a Markov logic network. The proposed system deals with partial segments, it is able to recover depth information even if the laser fails, and the integration is modeled through contextual information—characteristics not found on previous approaches. Experiments in pedestrian detection demonstrate the effectiveness of our method over data sets gathered in urban scenarios.

1. Introduction

The use of perception machines has been growing everyday, leading us to believe they will be mandatory in every transportation system in the near future. This way, multi-sensor fusion will play a crucial role in incorporating this technology in machines of daily-life use. Among different types of sensor combinations, rangefinders (e.g., laser scanners) and vision have been increasingly used in robotics. There are twofold reasons for that: The cost of both sensors has been decreasing significantly and the complementary nature of the integration brings acceptable performance in dynamic urban scenarios. Although these sensors can be used alone (vision [1] or laser [2]), the integration of them is usually necessary to cope with the diversity of outdoor situations.

So far, fusion of ranging and vision sensors has been performed by assuming that the probability to find an object is identical and usually independent, in both sensor spaces. These fusion approaches are based on laser region of interest (ROI), and can be accomplished in two ways: (1) sensor-driven classifiers (classifier-level fusion), or sensor-driven features (feature-level fusion) are integrated with independent and identically distributed (IID) assumptions, or (2) a laser scanner segmentation method is used to find the most likely ROIs; in the next step, these ROIs are projected into the image, being finally named by an image classification system. Though the goal of those architectures seems to be clear-cut to speed up image object detection, they may fail in many situations. Fig. 1 illustrates two scenarios where those two fusion approaches usually produce wrong outputs. In Fig. 1(a), the partial segment would be discarded by traditional fusion methods, since it represents segments below a certain established threshold. In other words, laser segments, in traditional methods, are expected to be full, ranging usually over lower and upper thresholds, in order to be considered as a hypothesized ROI. Fig. 1(b) shows a situation where only the vision system detected the occluded pedestrian. That pedestrian is also discarded by ROI-based fusion methods.

Contextual information [3,4] may be beneficial to overcome those issues, as we will demonstrate along this paper. We propose to integrate semantically both sensors in pedestrian detection tasks, contributing to improve the fusion of a laser scanner and a camera. Since we are focused on outdoor scenarios, we need to cope with the problems of shadow covered areas, illumination change and, mainly, object occlusion. To this end, our framework starts treating the laser segmentation problem, going from a coarse-to-fine segmentation to a segment labeling. It allows us to deal even with partial segmentation in our semantic framework. Pedestrian images are classified by a parts-based detector based on our ensemble of classifiers, proposed and extensively evaluated in [5]. The searching method exploits the contextual constraints given by the geometry of the laser sensing space. The spatial relationship between the attributes of objects is computed by means of a Markov logic network (MLN) [6]. Our main goal is to perform a frame-by-frame pedestrian detection, although the system can be adapted to any other object detection.

1.1. Previous work

Szavas et al. [7] entirely rely on laser ROIs in order to find probable areas where a pedestrian might be. Each image projected ROI is classified by a convolutional neural network (CNN). Broggi et al. [8] propose an on-spot pedestrian classification system,
which first uses a laser scanner to determine areas between vehicles, and then a Haar-like feature/adaboost classification system to name the projection of these areas in the image. The goal is to cope with the sudden appearance of pedestrians in urban traffic. Mählisch et al. [9] propose a spatio-temporal alignment to integrate a laser scanner and a monocular camera. For that, features in both sensor spaces are extracted and passed to a Bayesian classifier to detect cars. Douillard et al. [10] propose an approach similar to the previous one, but rather than a Bayesian rule they use a conditional random field (CRF). They are able not only to fuse laser and image features, but also to find a temporal relationship between features of sequential sensor readings. Premebida et al. [11] propose to classify features in laser space with a Gaussian mixture model (GMM) while Haar-like features are extracted over the image objects and classified by an adaboost. The confidence scores of the classifiers feed a Bayesian rule. Spinello et al. [12] propose a cooperative fusion of independent sensor detection. In laser space, an Adaboost is used to extract a set of boosted features, which posteriorly feed a CRF to classify pedestrians and cars. In image space, an implicit shape model (ISM) is applied in order to classify the objects in the scene. Both sensors are fused by means of a tracking system used in each sensor space with correspondent data association.

No aforementioned work uses contextual information to integrate both sensors, failing as laser does. In our work, we follow another direction, building a model from sensor raw data to meaningful parts, taking advantage of contextual information. This led us to conceive a system which can explore relevant characteristics of each sensor space in a semantic fusion framework.

1.2. Document organization

This article is organized as follows. In Section 2, the general idea of the framework is described. Section 3 introduces the steps to get the human-part labels from laser points. Section 4 recalls the idea of the framework is described. Section 3 introduces the steps to get the human-part labels from laser points. Section 4 recalls the idea of the framework is described. Section 3 introduces the steps to get the human-part labels from laser points. Section 4 recal...
In the last stage, all information is processed according to first-order formulas, subsequently structuring a Markov random field (MRF) with a grounded first-order formula in each node, providing marginal probabilities for each input.

3. From laser points to labels

The first step in object detection using laser is usually the extraction of geometric features [9–11]. Conversely, we follow a different direction based on a featureless approach. The goal is to cluster meaningful segments in a coarse-to-fine approach, labeling them with a level of confidence given by a template matching procedure. This latter one is based on Procrustes analysis [15].

3.1. Segmenting laser points

3.1.1. Coarse segmentation

Let \( C = \{c_i\} \), where \( z = 1, \ldots, Z \), be the number of points of a laser scan. After coarse segmentation, we have \( C = \{c_n\} \), where \( n = 1, \ldots, N \), coarse segments. The set \( C \) is obtained by convolving the points in \( C \) with a kernel mask \([-1,1]\), clustering those points which are at a distance of another point smaller than a threshold \( \Delta \). Only segments with size \( \epsilon [0.25, 1] \) m are considered. These values of segments range from partial segments to large human silhouettes. We assume the size of a segment as the Euclidean distance between its endpoints. Since we propose a featureless approach, this way of measuring a segment size will not influence the laser detection on distinguishing between flat and rounded objects or object parts. This distinction is done by the Procrustes analysis (superimposition on reference objects), which provides a scored identity of each body part.

3.1.2. Fine segmentation

Let us note that inside a coarse segment \( c_n \), there might be many holes (spaces between groups of points) coming from the scanning process when the pedestrian has body parts away from one another. This prevents any attempt to label \( c_n \) directly. To tackle this problem, we subsegment the set \( C \) into finer segments \( F = \{f_m\} \), where \( m = 1, \ldots, M \) due to the waist-level laser mounting (see Section 6.1, for details). The step of finding the set \( F \) is done by a \( \beta \)-skeleton method, a parameterized family of relative neighborhood graphs (RNG) proposed by Kirkpatrick and Radke [16], Toussaint [17] and Urquhart [18] initially established the theoretical foundations of RNGs to cluster planar points. \( \beta \)-skeleton RNGs represent then a parameterized evolved method of this family of graphs. Following, we present general definitions for \( \beta \)-skeletons:

**Definition 1.** Let \( V \subseteq \mathbb{R}^2 \) be a vertex set. For any pair \((p,q) \in V \times V\), the set \( A_{p,q}(\beta) = B_1((1-\beta/2)p + (\beta/2)q) \cap B_1((1-\beta/2)q + (\beta/2)p) \cap (\beta/2)dp(p,q) \cap B_1(-\beta/2)dp(p,q) \) is called a lune, and \( B(x,r) \) is a circle of radius \( r \) centered at \( x \).

**Definition 2.** The \( \beta \)-skeleton, \( G_\beta(V) \), with vertex set \( V \subseteq \mathbb{R}^2 \), is defined to be the graph on \( V \) with edge set \( E \) defined by lunes; thus \( p,q \in E(G_\beta(V)) \Leftrightarrow A_{p,q}(\beta) \cap V = \emptyset \).

It is noteworthy that: (1) if \( 0 < \beta < 1 \), \( A_{p,q}(\beta) \) contains no points other than \( p \) and \( q \); (2) if \( \beta = 1 \), the circle of diameter \( (p, q) \) contains no other points than \( p \) and \( q \); (3) if \( \beta > 1 \), the union of two circles having diameter \( \delta(p, q) \) contains no points other than \( p \) and \( q \). In the last case, \( G_\beta \) is not guaranteed to be connected. It is straightforward to conclude that the value of \( \beta \) controls the sparseness of the graph. Hence, we conceived a clustering tendency index, \( T \), to fit the best \( \beta \) to our goal. To this purpose, we expect that the vertices of \( G_\beta \) have degree equal to 1, and the number of clusters is not greater than 3. That index is defined as

\[
T = \log \left( \frac{1}{|C|} \sum_{i=1}^{|C|} (\gamma_i + \zeta_i) e_i + h_i \right),
\]

where \(|C|\) is the cardinality of the set \( C \); the functions \( \gamma_i, \zeta_i, e_i \) and \( h_i \) are given by

\[
\gamma_i = \frac{1}{K} \sum_{j=1}^{K} e_{ij},
\]

where \( K_i \) is the number of clusters in a coarse segment \( i \), and \( e_{ij} \) is the edge distance between two vertices with \( j \) indexing the number of edges inside a cluster \( k \) (fine segments).

\[
\zeta_i = \begin{cases} 
\frac{1}{K - 1} \sum_{k=1}^{K-1} S_k & \text{if } K_i > 1, \\
0 & \text{otherwise},
\end{cases}
\]

where \( S_k \) is the Euclidean distance between two fine segments.

\[
e_i = \begin{cases} 
1 & \text{if } \max(\deg(v_i)) = 0, \\
\max(\deg(v_i)) & \text{otherwise},
\end{cases}
\]
where $\text{deg}(v_{ij})$ is the number of edges incident on the vertex $v_{ij} \in V < C$.

$$h_i = \begin{cases} 1 & \text{if } K_i \leq 3, \\ K_i & \text{otherwise.} \end{cases}$$

(5)

The rationale behind function $T$ is to penalize values of $\beta$ which build random graphs whose degree of the vertices is other than 1, and also the built graphs are too sparse. Hence, the goal is to take the lowest value of $T$. For a set of laser scans, $\beta$-skeleton graphs are computed with $\beta$ varying in the interval [1, 3] (stride 0.1), averaged over all laser scans. Fig. 3(a) shows the results for the $\beta$ values averaged over a training data set. $\beta = 2.1$ was the chosen value (see Fig. 3(d), for an example).

Although this procedure gives a clustering tendency, it can still fail in the clustering boundaries. However, rather than subsequently pruning the graphs at this level, this is done in the labeling stage.

### 3.2. Labeling segments

This step starts with labeling each fine segment $f_k \in F$ as 4 different classes: “torso”, “arm”, “potential occlusion” and “noise”. “Noise” segments are always discarded. Each label is initially given by measuring the width of the segments. It is subsequently relaxed by thresholding the coarse segments. To overcome that, a relabeling process relies on the view of an inconsistency on the boundaries of the segments. In this stage, it is possible to find wrong labels because of the boundaries on thresholding the coarse segments. To overcome that, a relabeling procedure was conceived based on the rules on Table 1. The relabeling process relies on the view of an inconsistency on the semantic form of the grouping parts of a pedestrian silhouette. After all, final decisions will consider not only the label but also its confidence score.

To compute the confidence score of each label, a Procrustes analysis is performed, which matches the shape of a target segment to a hierarchy of reference shapes, after filter translation, scale and rotation information. Procrustes is a statistical analysis of shapes, proposed by Huerley and Cattel [15], commonly used in biology data to match landmark data [19]. Landmark data are corresponding points between two shapes, which are chosen manual or automatically. In our case, the landmarks are the laser points themselves. Next, necessary definitions [20] to perform Procrustes analysis are presented.

**Definition 3.** An $m \times m$ rotation matrix satisfies $I^T \Gamma = \Gamma I^T = I_m$ and $|\Gamma| = +1$. The set of all $m \times m$ rotation matrices is known as the special orthogonal group $SO(m)$.

**Definition 4.** The Euclidean similarity transformations of a configuration matrix $X$ are the set of translated, rotated and isotropically rescaled $X$, i.e.,

$$\{zXI + 10^T : z \in \mathbb{R}^+, \Gamma \in SO(m), \theta \in \mathbb{R}^m\},$$

(6)

where $z \in \mathbb{R}^+$ is the scale factor, $\Gamma$ is the rotation matrix and $\theta$ is a translation $m$-vector.

In the Procrustes analysis, the matrix $X$ denotes the set of the landmarks points (shape). To conform a target shape $X_1$ to a reference shape $X_2$, it is necessary to filter one transformation at a time. To filter the translation, centered landmarks, $X_c$, are obtained from the target landmarks $X$ according to

$$X_c = CX,$$

(7)

where $C$ is called centering matrix [21]. Multiplication by $C$ has the same effect as subtracting by the mean of $X$. $C$ is expressed by

$$C = \frac{1}{n} I_n \frac{1}{n},$$

where $I_n$ is the identity matrix of size $n$, $1$ is the column-vector of $n$ ones, and $T$ is the transpose.

In case of scale filtering, it is necessary to standardize for size, which is done as follows:

$$\phi = \alpha X_c,$$

(8)

where $\phi$ is invariant under translation and scaling of $X$, and is called the pre-shape of $X$, $\alpha = 1/\|X_c\|$, with $\|X_c\| = \sqrt{\text{trace}(X_c^T X_c)}$.

Then, the full Procrustes distance, $d_{\text{proc}}$, where $0 \leq d_{\text{proc}} \leq 1$, between $X_1$ and $X_2$ with pre-shapes $\Phi_1$ and $\Phi_2$, is given by

<table>
<thead>
<tr>
<th>Seq. of labels in a coarse segment</th>
<th>Relabeling to</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Potential”/“potential”</td>
<td>→ “Torso”</td>
</tr>
<tr>
<td>“Potential”/“torso”</td>
<td>→ “Torso”</td>
</tr>
<tr>
<td>“Arm”/“potential”/“potential”</td>
<td>→ “Arm”/“torso”</td>
</tr>
</tbody>
</table>

Table 1

Relabeling rules.

---

Fig. 3. Our final results of our clustering tendency index in (a) over a training data set, and examples of our graph-based fine segmentation over coarse segments of a human torso and hand. From (b) to (f), as $\beta$ grows, the sparseness of the random graphs increases.
minimizing over rotations and scale to find the closest Euclidean distance between \( \Phi_1 \) and \( \Phi_2 \), such as
\[
d_{\text{proc}}(X_1,X_2) = \left\{ 1 - \left( \sum_{i=1}^{R} \lambda_i \right)^{1/2} \right\}^{1/2},
\]
where \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{R-1} \geq |\lambda_R| \) are the square roots of the eigenvalues of \( \Phi_1^T \Phi_2 \Phi_2^T \Phi_1 \), which should be sorted with respect to their absolute values.

The rotation that minimizes the distance between \( \Phi_1 \) and \( \Phi_2 \) is given by
\[
\hat{R} = UV^T,
\]
where \( U,V \in \text{SO}(m) \) and \( \Phi_1^T \Phi_1 = VAV^T \), with \( A = \text{diag}(\lambda_i) \).

The scale that minimizes the distance between \( \Phi_1 \) and \( \Phi_2 \) is
\[
\hat{\beta} = \frac{1}{R} \sum_{i=1}^{R} \lambda_i.
\]

To match the target shapes to the reference shapes, a tree of laser segments were collected from persons with different silhouettes. The pre-shape of the collected segments were averaged over 10 samples. The tree of reference landmarks are comprised of sided, back and front shapes. Back and front models were used with hands close and far from the side of the body, while side models were considered in left and right positions.

Procrustes analysis imposes equality of the two sets of landmarks. As equality is not guaranteed, that constraint is overcome by extrapolating or shrinking points in the target set. Since the shape of a pedestrian captured by the laser tends to hold a U-form, a parabolic regression suffices to obtain extra points, satisfactorily. On the other hand, if the cardinality of the target set of landmarks is smaller than the cardinality of the reference set, the last points are discarded. The Procrustes distance for each segment is given by the maximum value over the distances found between the target landmarks and all the landmarks in the tree. All steps to segmenting and labeling the laser points are described in Algorithm 1. Some results of the algorithm over a testing data set are shown in Fig. 4.

Algorithm 1. Laser segmenting and labeling.

```plaintext
input: \( \{l_i\}_{i=1}^n \) points from one scan and \( \{x_{rj}\}_{j=1}^m \) reference landmarks
output: A set of pairs \( \{f_m,d_m\}_{m=1}^M \subset F \), Dists
1 Dists \( \leftarrow \) initialize an array of Procrustes distances;
2 \( \{c_n\} \leftarrow \) CoarseSegment (\( \{l_i\} \));
3 \( \{f_m\} \leftarrow \) FineSegment(\( \{c_n\} \));
4 NeedRelabel \( \leftarrow \) Label(\( \{f_m\} \));
5 if NeedRelabel then Relabel (\( \{f_m\} \));
6 for each \( f_m \) do
7 \( \text{for each} \ (x_{rj}) \subseteq \{x_{rj}\} \) do
8 \( \text{if} \ |l_i| > |x_{rj}| \text{ then} \) Extrapolate(\( l_i \));
9 \( \text{else if} \ |l_i| < |x_{rj}| \text{ then} \) Shrink (\( l_i \));
10 Dists \( \leftarrow \) Match (\( l_i \),\( x_{rj} \));
11 end
12 end
13 Dists \( \leftarrow \) max(Dists);
```

4. Image object classifier

4.1. Searching objects in laser sensing space

A typical image object detector uses a sliding window method to find a pedestrian in all image scales and positions. Instead, we
explore here the 3D geometry given by the laser sensing space. There are twofold reasons for that: (1) even if the laser classification fails, we are able to recover approximate depth information and (2) by constraining the detection to meaningful places, we can avoid many false alarms.

Let \( p_1 = (0, -1.0, \text{dist}_1)^T \) and \( p_2 = (0, 1.0, \text{dist}_2)^T \) be two planes in homogeneous coordinates given with respect to laser reference frame, representing the ground and a plane which passes throughout the camera, respectively (see Fig. 9). Both planes are parallel to the laser plane. This way, \( \text{dist}_1 \) (0.9 m) represents the distance of the laser to the ground, and \( \text{dist}_2 \) is the distance between the center of the laser and the center of the camera lens (\( \text{dist}_1 = \text{dist}_2 \)). A sliding window in laser reference frame is slid in the viewport formed of \( p_1 \) and \( p_2 \) as the top and bottom bounding planes (see Fig. 5). This window of size 1.0 m \( \times \) 1.8 m is shifted onto horizontal and vertical directions in laser sensing space, ranging over 2 m up to 20 m, with a stride of 0.20 m. Only the projected windows lying on the image frame are used, after being projected with the method proposed in [14]. For each window, the depth information is kept for posterior reasoning (estimated depth information when the laser fails).

### 4.2. Image classification

The image classification system is based on our proposed ensemble of classifiers called HLSM-FINT, found in [5], which stands for histogram of oriented gradients (HOG), local receptive features (LRF), support vector machines (SVM), multi-layer perceptron (MLP) and fuzzy integral (FINT). The features HOG and LRF (this latter obtained by a convolutional neural network) are classified by the classifiers SVM and MLP, whose outputs feed a FINT-based fusion module. The main idea of the method is to increase the synergism of the components by means of the fuzzy integral. The choice of the feature representations relies on the characteristic of relative lighting and shift invariance. It was also experimentally demonstrated in [5] that HOG is better to classify pedestrians, whilst LRF is better to classify non-pedestrians. HLSM-FINT was in-depth evaluated over DaimlerChrysler and home-grown data sets, presenting a consistently better performance among other methods and the component classifiers.

HLSM-FINT was originally used monolithically. Here, we used the data sets of Institut National de Recherche en Informatique et Automatique (INRIA) [22] to train a parts-based detector over two parts of the human body (see Fig. 6(a)) with the same HLSM-FINT parameters found in [5], and with a training data set distribution of 2416 pedestrians and 15,000 non-pedestrians (for upper and lower parts). This choice avoids parts not being properly detected as it may happen in limbs-based approaches, in far distances. The 54 \( \times \) 108 pixel reference image window has an upper body height of 34 pixels, starting at position (0,0), and a lower body height of 24 pixels, starting at location (0,48), in reference to the origin of the detection window. After applying the parts-based HLSM-FINT in each of the projected sliding windows, an object can be bounded by many windows. Hence, a non-maxima suppression algorithm is used in the image window projections to prune those extra windows, keeping the depth information. Some image samples classified by our detector are illustrated in Fig. 6(b).

### 5. Semantically integrating the parts

In our framework, we begin to model the interdependency of the data by extracting meaningful information from object parts. Later, the spatial relationship of pedestrian attributes is embodied in a first-order knowledge base (KB), which is subsequently modeled in a MRF, inside an MLN framework.
5.1. Markov logic network: foundations

A first-order KB is composed of a set of formulas. In turn, formulas can be formed by three types of symbols: constants, variables and predicates. Constants represent an object in the universe of interest. Variables range over objects. Predicates represent relationships among objects, or designate attributes to objects. Constants, variables and predicates are terms of the formulas. An interpretation assigns a symbol to each object in the universe of interest. First-order logic also makes use of quantifiers (e.g., universal quantifier, \( \forall \), existential quantifier, \( \exists \)) to define the domain of a variable. An atomic formula or atom is a predicate applied to a tuple of terms. A ground atom is a predicate which contains constants assigned by possible variables.

In first-order logic, if a world violates one formula, then it is impossible to occur. The first idea of Markov logic is to soften these constraints, that is, when a world violates a formula, it becomes less probable, but not impossible. To this end, each formula in a KB has a weight which represents how strong a formula is for a world. The higher the weight is, the greater the probability for a world to satisfy that formula. The main goal of MLN is thus to unify the fundamental advantages of first-order logic and MRF, dealing at the same time with complexity and uncertainty.

Definition 5. A first-order MLN is a set of pairs \((Q_i, w_i)\), where \(Q_i\) is a formula in first-order logic and \(w_i\) is a real-number weight. Each \(Q_i\) is a node of a MRF, while the relationship of each formula structures the way that each MRF is comprised.

Given different constants, a first-order MLN will produce different MRNs, which are called ground MLNs. Each ground MLN varies in size but keeping regularities in structure and parameters, given by the first-order MLN. The ground MLN is defined as

\[
P(X = x) = \frac{1}{Z_{\text{part}}} \exp \left( \sum_i w_i \eta_i(x_i) \right),
\]

where \(Z_{\text{part}} = \sum_{x \in X} \exp(\sum_i w_i \eta_i(x_i))\) is a partition function, \(\eta_i(x)\) is the number of true groundings of \(Q_i\) in \(x_i\), which is, in turn, the state of the \(i\)th atom in \(Q_i\).

The weights of an MLN can be learnt or hand-crafted. Weight learning can be performed generative or discriminatively. We used a discriminative approach, based on voted perceptron weighted satisfiability solver, which is demonstrated to outperform generative approaches [23]. Discriminative learning is performed by using a maximum a posteriori (MAP) inference over the gradients of the conditional log-likelihood, given by

\[
\frac{\partial}{\partial w_i} \log P_{\text{post}}(\psi | e) = \eta_i(\psi, e) - E_{\text{wa}}[\eta_i(\psi, e)],
\]

where \(\psi\) is a query predicate, \(e\) is an evidence predicate, \(E_{\text{wa}}[\eta_i(\psi, e)]\) is the expectation over the number of true groundings of formula \(i\) according to the MLN. A training data set was used to learn the weights, \(w_i\), of our MLN (refer to Section 6). After training an MLN, inference is performed by a combination of Markov chain Monte Carlo (MCMC) and WalkSAT, called MC-SAT [24].

5.2. Dealing with occlusion by spatial relationship between laser segments

In the laser space, the occlusion problem is tackled by a set of situation assertions defined from the geometry of the scene. These assertions rely inherently on partial segmentation. As a result, if a partial segment is “near” by other segment, it is a candidate to be considered occluded. The concept of “nearness” is used over partial segments according to the angle between the end point of a segment and the start point of its closer segment (Fig. 7). In practice, an object is close to another object if it is not greater than 5 times the laser angular resolution. This situation indicates initially a high hypothesis of the object to be occluded by a nearest object in its front. A z-buffer analysis is therefore made for each object and its vicinity.

In this stage, we extend the concept of partial segmentation not only to potential occlusion segments but also to those segments which present only one arm (see Fig. 8(b)). The relationship of the attributes as well as their “nearness” is modeled as first-order logic KB (formulas 7 and 8 in Table 2). Table 2 summarizes the first-order formulas used in our KB for the proposed semantic fusion. An object “\( o \)” can be either a window in image space or a segment (also surrounded by a bounding box in the image) in laser space, represented by a sequential number as objects are detected.

Some results of occlusion detection are illustrated in Fig. 8. In Figs. 8(a) and (b), some results of the application of formulas 7 and 8 are illustrated, while Fig. 8(c) shows two cases where potential occlusions are discarded.

5.3. Semantic multi-sensor fusion

After modeling occlusion in laser space, formula 9 (see Table 2) describes the problem of occlusion in the image space. Actually, this formula also works as the kernel of the fusion between laser and camera by logically adding it as a disjunction to the formulas 10 to 15. In other words, if for a given object \( o \), this object is only seen in the image (upper or lower body), formula 9 will be considered alone, otherwise if a laser window overlaps an image window, it will be responsible to increase the weight of formulas 10–15, in the training stage. If we find a torso as a unique part (formula 10), that is, there was just one part \( f_m \), then a high confidence score (low Procrustes distance) is necessary to validate it as a pedestrian. Similarly, when a coarse segment, \( c_m \), is formed by a potential occlusion (formulas 11 and 12), it means that the segment could have not been labeled appropriately due to the limit boundaries in the coarse step. In this case, the score of the coarse segment provides a hint to establish it as a pedestrian or not. The last formula represents the highest evidence to assign a coarse segment to a pedestrian, because the segmentation was...
It is noteworthy that there are different thresholds for “HighScore” (formulas 10 to 14) according to the body parts being analyzed. These values represent established thresholds for pedestrians comprised of different sets of parts (e.g., “arm”/“torso”, “arm”/“potential”).

6. Experimental validation

In this section, we describe the methodology used to validate the proposed semantic fusion with experiments in a real urban scenario.

6.1. Experimental setup

The vehicle sketch as well as some pictures of the used sensors are illustrated in Fig. 9. The laser scanner was mounted at a height of 0.9 m in order to detect pedestrians at the waist level. The system sensing ranges from 2 up to 20 m, providing sufficient time to stop the low-speed vehicle (the maximum speed of the vehicle is approximately 30 km/h) whenever a pedestrian is detected in critical areas. The laser scanner was set with an aperture of 100° and angular resolution of 0.25°, while the camera forms a field of view of 45°. For each sensor, it is assigned a thread in a dual-core computing system with CPU affinity and synchronization barriers. Sensor data was acquired, and kept entirely in the main memory during acquisition process, being saved to the hard disk only at the end of the acquisition trajectory. These procedures were performed to guarantee synchronization of the two sensor data, acquired at a rate of 15 fps.

Each image was manually annotated by using a special software conceived to create the ground truth. The laser detection is also evaluated in the image space, after (coarse) person segments being properly projected and evaluated by MLN. The annotations included all visible objects in the images, even partially occluded, considering only objects in the range of 2 up to 20 m. The maximum effective speed to collect the data sets was 9.7 km/h. The vehicle was driven through our university campus in a crowded area, with many parked and moving cars. Besides persons naturally walking on the campus, some other persons were intentionally crossing in front of the vehicle in order to cast occlusion situations. Table 3 summarizes the characteristics of the collected data sets.

6.2. Evaluation of the proposed method

The collected data sets were used to train and test the proposed method. Sequence #1 was used to train the MLN, and to select the clustering tendency index, while Seq. #2 was used to test the proposed methods. The sequences were gathered from distinct situations (different days and different objects). The detection performance was assessed by receiver operating characteristics (ROC) curves. To build the MLN system, we have used the Alchemy library.

The ROC curves were plotted over the Seq. #2 for the classification methods in each sensor space, as well as for the proposed semantic fusion (see Fig. 10). The operating points of all curves were chosen to be at 0.5 false alarm per frame (FAPF), aiming at maximizing the hit rate (HR), while minimizing the FAPF. This operating point works as the one to show a spot

---

2 Data sets, annotations and transformation matrix are available in http://www.isr.uc.pt/~lreboucas.
3 Available in http://alchemy.cs.washington.edu/
comparison among the various plots (boxed at the legend of the graphs). In the laser space, we compared our segmentation method with an iterative end point fit (IEPF) segmentation, proposed by [25] and surveyed in [26] as the best method amongst others for an indoor laser segmentation task. Both segmentation methods were classified by our Procrustes analysis (PA) along with our reference tree template matching approach (see Figs. 10(a) and (b)). For that, we subsampled our laser data set, initially acquired with a laser angular resolution of 0.25°, to get extra resolutions of 0.50°, 0.75° and 1°. In our method, the points were segmented with $D = 0.25\,\text{m}$, in the coarse segmentation step, being noise segments discarded. In average, the HR of the detection method using our segmentation method is 10.5 percentage points better. Concerning the image classification, Fig. 10(c) shows the results for monolithic (53% of HR) and parts-based HLSM-FINT (63% of HR). Although our detector reached high performance in [5], here it was challenged with a more difficult data set, with a considerable number of occlusions and

### Table 2

First-order formulas used in the semantic fusion.

<table>
<thead>
<tr>
<th>No.</th>
<th>First-order logic</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$o$, Person($)</td>
<td>Query over object $o$</td>
</tr>
<tr>
<td>2</td>
<td>$o \ni \forall p$, Part(id,t,p)</td>
<td>Object $o$ with part of type $t$, in position $p$</td>
</tr>
<tr>
<td>3</td>
<td>$o \ni \forall t$, Unique_Part(o,t)</td>
<td>Object $o$ of type $t$</td>
</tr>
<tr>
<td>4</td>
<td>$o$, Occluded($)</td>
<td>Object $o$ is occluded by its nearest object</td>
</tr>
<tr>
<td>5</td>
<td>$o$, HighScore($)</td>
<td>Object $o$ with small Procrustes distance</td>
</tr>
<tr>
<td>6</td>
<td>$o$, View($)</td>
<td>Object $o$ with a body part $k$</td>
</tr>
<tr>
<td>7</td>
<td>$o$, Unique_Part(o,‘‘potential’’) $\land$ Occluded($) $\Rightarrow$ Person($)</td>
<td>If an object $o$ has any potential occlusion segment and it is potentially occluded by its nearest object, it is a pedestrian</td>
</tr>
<tr>
<td>8</td>
<td>$o$, Unique_Part(o,‘‘arm’’) $\land$ Occluded($) $\Rightarrow$ Person($)</td>
<td>If an object $o$ has only one arm segment and it is potentially occluded by its nearest object, it is a pedestrian</td>
</tr>
<tr>
<td>9</td>
<td>$o$, View(o,‘‘upper’’) $\lor$ View(o,‘‘lower’’) $\Rightarrow$ Person($)</td>
<td>If an upper or a lower part of an object $o$ is viewed, it is a pedestrian</td>
</tr>
<tr>
<td>10</td>
<td>$o$, Part(o,‘‘arm’’,i) $\land$ Part(o,‘‘torso’’,j) $\land$ HighScore(o) $\Rightarrow$ Person(o)</td>
<td>If an object $o$ has arm and torso with Proc. distance smaller than 0.25, it is a pedestrian</td>
</tr>
<tr>
<td>11</td>
<td>$o$, Part(o,‘‘arm’’,i) $\land$ Part(o,‘‘arm’’) $\land$ HighScore(o) $\Rightarrow$ Person(o)</td>
<td>If an object $o$ has arm and arm with Procrustes distance smaller than 0.10, it is a pedestrian</td>
</tr>
<tr>
<td>12</td>
<td>$o$, Unique_Part(o,‘‘torso’’) $\land$ HighScore(o) $\Rightarrow$ Person(o)</td>
<td>If an object $o$ has only one torso and a Procrustes distance smaller than 0.10, it is a pedestrian</td>
</tr>
<tr>
<td>13</td>
<td>$o \ni \forall j$, Part(o,‘‘arm’’,i) $\land$ Part(o,‘‘potential’’,j) $\land$ HighScore(o) $\Rightarrow$ Person(o)</td>
<td>If an object $o$ has two arms and one potential, and it has a Procrustes distance smaller than 0.20, it is a pedestrian</td>
</tr>
<tr>
<td>14</td>
<td>$o \ni \forall j$, Part(o,‘‘arm’’,i) $\land$ Part(o,‘‘potential’’,j) $\land$ Part(o,‘‘arm’’,i) $\land$ HighScore(o) $\Rightarrow$ Person(o)</td>
<td>If an object $o$ has 2 arms and one torso, then it is a pedestrian</td>
</tr>
</tbody>
</table>

### Table 3

Characteristics of collected data sets.

<table>
<thead>
<tr>
<th>Sequence</th>
<th># Frames</th>
<th># Annotated objects</th>
<th>Dist</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1 (train)</td>
<td>2672</td>
<td>3429</td>
<td>160 m</td>
</tr>
<tr>
<td>#2 (test)</td>
<td>2157</td>
<td>3692</td>
<td>160 m</td>
</tr>
</tbody>
</table>
real urban situations. The results show that parts-based HLSM-FINT gives better performance. Finally, Fig. 10(d) shows that, in laser detection, 71% of HR were found for segmentation and classification with formulas 7 and 8; the proposed method (with all MLN formulas) achieved 80.8% of HR, increasing the performance of our parts-based detector by 17.8%, and the performance of the MLN-based classification by 13.8%.

Over Seq. #2, the detected objects were grouped according to the distances from the laser as well as vehicle speed in order to study the robustness of our method. The maximum distance where an object was annotated in Seq. #2 was roughly 16 m. In Table 4, the HR and false alarm rate (FAR) were computed in where an object was annotated in Seq. #2 was roughly 16 m. In study the robustness of our method. The maximum distance performance of the MLN-based classification by 13.8%.

performance of the proposed method performance. It is worthy recalling that our method is meant to be employed in a low-speed vehicle. According to the results, the increase of the vehicle speed turns the performance of the method to drop in average only 3.75 percentage points to each speed range illustrated in the table.

Some results of synergistic detection in laser and image spaces are shown in Fig. 11. The first line illustrates the detectors in each sensor space, while the second line shows the fusion result. When only the output of the image detector is chosen by the fusion method, then an estimated distance from the vehicle is given. In the illustrated scene, it can be observed that, in the two left-most images, the laser segmentation failed most of the times on “Pedestrian A”, who was crossing the street in front of the group of persons, providing only few points due to the low reflectivity of the black shirt in a far distance. This problem is inherent to the laser scanner used, and it can be tackled with a more accurate sensor. In this case, our image detector was effective to improve the accuracy of the recognition.

7. Conclusions and future work

A novel approach to integrate laser and vision data has been presented in this paper. The system is based on a semantic fusion of parts-based detectors in laser and image spaces. Our system overcomes limitations found in existing fusion methods, providing characteristics such as: treating partial segmentation, recovering depth information even when the laser segmentation fails, and a

<table>
<thead>
<tr>
<th>Vehicle speed (m)</th>
<th>Vision detection</th>
<th>Laser detection</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>2–6</td>
<td>HR = 40%/FAR = 0.5%</td>
<td>HR = 80%/FAR = 1%</td>
<td>HR = 82%/FAR = 1.8%</td>
</tr>
<tr>
<td>6–10</td>
<td>HR = 65%/FAR = 2%</td>
<td>HR = 85%/FAR = 1.8%</td>
<td>HR = 90%/FAR = 2.5%</td>
</tr>
<tr>
<td>10–14</td>
<td>HR = 60.2%/FAR = 1.5%</td>
<td>HR = 70.4%/FAR = 2.3%</td>
<td>HR = 80%/FAR = 3.5%</td>
</tr>
<tr>
<td>14–16</td>
<td>HR = 30%/FAR = 5%</td>
<td>HR = 45%/FAR = 3%</td>
<td>HR = 50.4%/FAR = 4.5%</td>
</tr>
</tbody>
</table>
fusion method based on non-IID assumptions. In laser space, a featureless classification approach was proposed by using semantic information. The image search was done by projections of sliding windows slid in laser sensing space. This was particularly important in recovering the depth information when the laser failed; yet, a two parts-based detector aided in dealing with occlusions in image space. An MLN provided a comprehensive framework to perform the final fusion. Our proposed system increased the worst individual detection by 17.8 percentage points, reaching 80.8% of HR, with 0.5 of FAPF, in our gathered test data set. These results demonstrate how effective is the proposed method for sensing real urban situations in low-speed vehicles.

Our future research will follow three directions: Building a reliable tracking system not only to reduce false alarms provided by frame-by-frame detection, but also to estimate the trajectory of pedestrians with indications of their behaviors; parallelization of the algorithm to run on-the-fly, and incorporation of knowledge about the navigation to include context-awareness in specific places of the vehicle trajectory.

Acknowledgements

This work was supported by Fundação para a Ciência e Tecnologia de Portugal (FCT), under Grant PTDC/EEA-ACR/72226/2006. Luciano Oliveira was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), programme of Ministry of Education of Brazil, scholarship no. BEX 4000-05-6.

References


Table 5

<table>
<thead>
<tr>
<th>Vehicle speed (km/h)</th>
<th>Vision detection</th>
<th>Laser detection</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR = 70%/FAR = 2%</td>
<td>HR = 85%/FAR = 1%</td>
<td>HR = 91.3%/FAR = 2.8%</td>
</tr>
<tr>
<td>0–3</td>
<td>HR = 68.1%/FAR = 2.8%</td>
<td>HR = 82.7%/FAR = 3.1%</td>
<td>HR = 87%/FAR = 3.5%</td>
</tr>
<tr>
<td>6–9.7</td>
<td>HR = 65.6%/FAR = 2.9%</td>
<td>HR = 79.1%/FAR = 4.9%</td>
<td>HR = 83.8%/FAR = 5%</td>
</tr>
</tbody>
</table>

Fig. 11. An example of crowded scene where there is a perfect cooperation between vision (lighter bounding box, white) and laser (darker bounding box, blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

About the Author—LUCIANO OLIVEIRA received his B.CS. and M.Sc. in Mechatronics degrees from Federal University of Bahia, Brazil, in 1997 and 2005, respectively. Since 2006, he is a Ph.D. student at Institute of Systems and Robotics, at University of Coimbra, Portugal. In 2007, he was awarded as a winner of NiSIS Competition, Problem Task: “Analysis and Classification of the DaimlerChrysler Automotive Dataset Images”.

About the Author—URBANO NUNES received the B.Sc. and Ph.D degrees in Electrical Engineering from the University of Coimbra, Portugal, in 1983 and 1995, respectively. He is currently an Associate Professor at the Department of Electrical and Computer Engineering of the University of Coimbra. He is the Coordinator of the Automation and Mobile Robotics Group and the Coordinator of the Mechatronics Laboratory. He has been responsible for several funded projects in the areas of mobile robotics and intelligent vehicles. He was recipient of the IEEE ITS Society Outstanding Service Award in 2006.

About the Author—PAULO PEIXOTO received his B.Sc., M.Sc. and Ph.D. degrees in Electrical Engineering from the University of Coimbra in 1989, 1995 and 2003, respectively. He is currently an Assistant Professor at the Department of Electrical and Computer Engineering of the University of Coimbra and a Researcher of the Institute of Systems and Robotics, University of Coimbra. He has been responsible for several funded projects in the areas of computer vision, visual surveillance and intelligent vehicles.

About the Author—MARCOS SILVA received his B.Sc. from Industrial Electronics Engineering at Minho University, in 1998, and M.Sc. degrees in Electrical Engineering from the University of Coimbra in 2005. He is a Ph.D. student at Institute of Systems and Robotics, at University of Coimbra, Portugal, researching in the field of Cybercar systems. His main interests reside in microprocessors and robotics. He is an Assistant Professor at Coimbra Institute of Engineering, since 2000.

About the Author—FERNANDO MOITA received his B.Sc. and M.Sc. degrees in Electrical Engineering from the University of Coimbra in 1992 and 2003, respectively. He is a Ph.D. student at Institute of Systems and Robotics, at University of Coimbra, Portugal. His main interest reside in sensors, signal processing and robotics. He is an Assistant Professor at Coimbra Institute of Engineering, since 1996.