Sensors **2014**, *14*, 18670-18700; doi:10.3390/s141018670

OPEN ACCESS

sensors

ISSN 1424-8220

www.mdpi.com/journal/sensors

Article

A Context-Aware-Based Audio Guidance System for Blind People Using a Multimodal Profile Model

Qing Lin and Youngjoon Han *

Electronic Engineering Department, Soongsil University, 511 Sangdo-Dong, Dongjak-Gu, Seoul 156-743, Korea; E-Mail: lqsdust@163.com

***** Author to whom correspondence should be addressed; E-Mail: young@ssu.ac.kr; Tel.: +82-02-820-0699.

External Editor: Vittorio M.N. Passaro

Received: 4 August 2014; in revised form: 15 September 2014 / Accepted: 18 September 2014 / Published: 9 October 2014

Abstract: A wearable guidance system is designed to provide context-dependent guidance messages to blind people while they traverse local pathways. The system is composed of three parts: moving scene analysis, walking context estimation and audio message delivery. The combination of a downward-pointing laser scanner and a camera is used to solve the challenging problem of moving scene analysis. By integrating laser data profiles and image edge profiles, a multimodal profile model is constructed to estimate jointly the ground plane, object locations and object types, by using a Bayesian network. The outputs of the moving scene analysis are further employed to estimate the walking context, which is defined as a fuzzy safety level that is inferred through a fuzzy logic model. Depending on the estimated walking context, the audio messages that best suit the current context are delivered to the user in a flexible manner. The proposed system is tested under various local pathway scenes, and the results confirm its efficiency in assisting blind people to attain autonomous mobility.

Keywords: electronic mobility aids; sensor fusion; object detection; Bayesian network; context-aware guidance; multimodal information transformation

1. Introduction

According to recent statistics [1], 285 million people worldwide are estimated to be visually impaired worldwide, and among these, 39 million are completely blind. The loss of independent mobility is a large problem for these blind people, and they have to rely on the white cane as their primary mobility tool. However, the white cane has a restricted searching range. Therefore, considerable efforts have been made over the last 20 years to complement white canes with various types of electronic devices to aid mobility. Compared with white canes, these devices can help monitor a wider range of the environment and provide helpful feedback in various modalities. The most critical aspects of developing these electronic mobility aid devices are two-fold: how to sense the environment and how to inform the blind user [2]. In general, the environment can be sensed with various sensors, such as ultrasonic sensors, laser sensors and cameras. Users can be informed via auditory or tactile sense.

In recent years, the camera has gained attention in studies on environment sensing because of its several advantages, such as providing a large sensing area and rich information. A single camera or stereo cameras have been widely used in building electronic mobility aid systems. A single camera is more compact and easier to maintain than stereo cameras are. However, it is difficult to recover depth from a single image. Therefore, to distinguish foreground objects from the background, systems using the single camera have to make strong assumptions about scene geometry and features of appearance. For example, the "NAVI (Navigation Assistance for Visually Impaired)" system [3] uses gray-scale features to discriminate objects from the background. It classifies image pixels into background or objects using a fuzzy neural network. In [4], a color histogram is used to discriminate the ground from the objects. In [5,6], diagonally distributed road boundaries are extracted to find the path area, and then, the objects inside path area are detected by quasi-vertical edges or changes in texture patterns. In [7], by mapping the original image to a top-view plane, objects are detected and classified in the top-view space using the geometric features of object edges. Despite the efforts made to detect objects from a single image without direct depth cues, the appearance and geometry models used in these systems are valid only in limited scenarios.

Compared to a single camera, stereo cameras are more popularly used for building mobility aid systems, because depth can be computed directly from pairs of stereo images. The dense depth map can be used as a valuable asset in object detection and scene interpretation. Some of the systems directly quantize a depth map into a rectangular block representation, and then convert it into tactile vibrations or 3D sound that is perceived by blind users. For instance, the "TVS (Tactile Vision System)" [8] and "Tyflos" system [9–12] convert quantized depth maps into vibrations on a tactile sensor array that is attached to the user's abdomen. The "ENVS (Electron-Neural Vision System)" [13] transforms a vertically divided depth map into electrical pulses that stimulate the user's fingers. In [14,15], a depth map is mapped into a virtual acoustic space.

Instead of mapping the depth map directly to other modalities, some other systems have attempted to improve the resolution of scene interpretation using the depth map. For example, in [16], 3D scene points recovered from the depth map are classified as either in the ground or the object by estimating the ground plane. A polar accumulative grid is then built to represent the scene. In [17], a multi-level surface patch model is built for object representation. In [18], object regions are extracted by segmenting them on a saliency map, and the corresponding depth is computed by using the depth map.

Moreover, some systems also include the SLAM algorithm to maintain a local 3D map for object detection, as in [17]. In [19], aerial objects are detected using a local 3D map built from a SLAM algorithm. Although many stereo-vision-based systems have attempted to improve the resolution of scene interpretation, inherent problems still exist in stereo-vision methods. First, the stereo-matching algorithm often fails in texture-less areas, where depth cues are largely unavailable. Second, the accuracy of the depth map is sensitive to illumination and artifacts in the scene. Noise contained in the depth map complicates the identification of low-level objects, such as road curbs. Moreover, the generation of an accurate and dense depth map is still expensive, which reduces the possibility of adding other useful functions with respect to limited computation resources on a portable platform.

Compared with stereo-cameras, laser sensors can produce accurate range data that is not easily affected by environmental conditions. However, this usually requires laser sensors to scan the space to get a full 3D map, which is inconvenient in a walking-guidance task. One such example is the hand-held point-laser device designed by Manduchi, R. [20]. This device requires the user to swing it around in the space for range data collection. Because of this constraint, in some other human navigation systems, laser sensors are used mostly for indoor localization and 2D map building instead of object detection. For example, a combination of a laser scanner and a gyroscope is used in an indoor localization tool [21]. The tool is designed to locate blind users inside a building by matching laser corner features with the corners recorded in a known building map. In [22], a laser scanner and an inertial sensor are fixed on a helmet worn by the user for the purpose of indoor-map building and self-localization. A similar application of a laser scanner is in a wearable multi-sensor system developed in [23], which enables multi-floor indoor mapping and localization. These systems are designed mainly for indoor-map building, and forward-looking laser sensors can only detect high-level objects that are higher than the mounting position.

Moreover, following the recent development of the RGB-D camera, guidance systems using this kind of camera have also emerged [24–28]. The RGB-D camera can obtain both an RGB color map and a depth map of the whole scene in real time. Therefore, it may be used conveniently in both object detection and scene interpretation. However, the RGB-D camera depends on emitted infrared rays to generate a depth map, and in an outdoor environment, the infrared rays can be easily affected by sunlight. Therefore, guidance systems developed using the RGB-D camera can only be used in indoor environments, which limits the range of its use in a mobility aid system.

In this paper, a context-aware audio guidance system is proposed. The contribution of this paper is three-fold. First, a downward-pointing multi-sensor configuration is proposed to cover all types of objects, especially low-level objects that are close to the ground. Second, a multimodal profile model is proposed to interpret the scene at a high resolution. The third contribution is an audio feedback scheme based on the walking context.

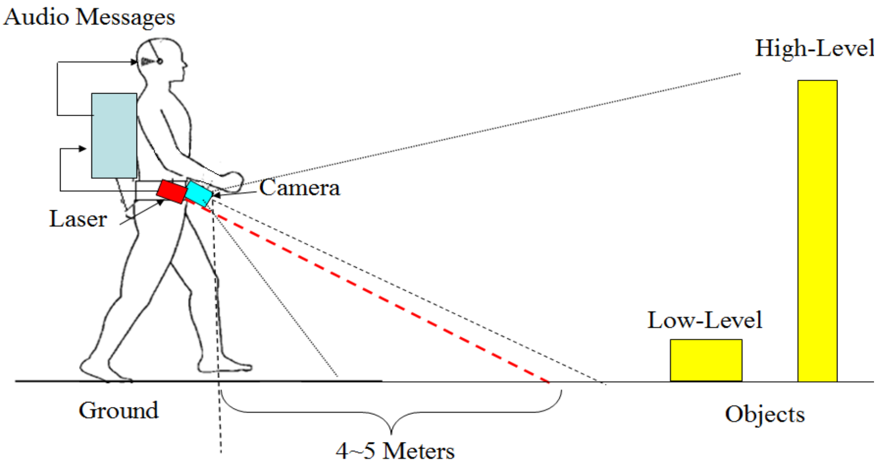
2. System Overview

The above review of the existing research revealed aspects that need improvement. For environment sensing, the improvement of scene interpretation resolution will help improve users' perceptions of the scene. An ideal scene interpretation resolution would include identifying the location of both the ground each object, as well as the type of each object. However, most existing systems can only interpret the scene at a level of two categories, either the ground plane or the object. They lack the ability to identify one object from the other. For user feedback, with the increase in scene interpretation resolution and context information, there will be an increased demand for delivering highly semantic information to the user. However, the commonly used tactile and virtual sound interfaces have difficulty in meeting this demand. In this paper, we propose a guidance system solution that aims at solving these problems.

The discussion in Section 1 indicates that laser sensors and cameras have properties that may allow them to complement each other. For example, the camera can capture an entire scene with rich intensity data in one frame shot, but it is hard to obtain stable and accurate depth cues from a single image. Although the laser sensor can provide stable and accurate range data, it requires a time-consuming scan to collect sufficient data that covers the whole scene. Therefore, these two types of sensors have the potential to be integrated in order to provide real-time multimodal data for use in the guidance tasks of blind people.

Figure 1 shows an illustration of the system prototype. A camera and a laser range finder are fixed with a downward-pointing angle on the waist of a blind user. In this viewing angle, both low and high objects can be captured by the combined sensors. As the user traverses a local environment, two sensors collect environment data in their respective modalities, and the incoming sensor data is further processed on a portable computer, where high-level information regarding scene geometry, objects and walking context is inferred. Finally, the inferred results are forwarded to the user in a flexible manner via audio messages.

Figure 1. Overview of the system prototype.



The system functions are divided into three parts: moving scene analysis, walking context estimation and audio message generation. The task of moving scene analysis is to detect the ground and objects, while estimating a safe path. Based on the output of moving scene analysis, the walking context estimation evaluates the current walking status as safe, normal or dangerous. Finally, according to the estimated walking context, the audio message generation selects critical information from the moving scene analysis and delivers it to the user at the right moment.

3. Moving Scene Analysis

3.1. Multimodal Profile Model

In real scene analysis, because of various ground conditions, sensor data noise and unstable sensor motion, large uncertainties exist when separating low-level objects from the ground in the sensor data. Similarly, because of large variances in the shape, size and appearance of objects, many uncertainties also occur when identifying individual objects. To minimize the uncertainties and to obtain the optimal inference of the scene content, a multimodal profile model based on a Bayesian network is proposed. The optimal scene model parameters are inferred by maximizing their joint probability distributions.

As illustrated in Figure 2a, the proposed multimodal profile model is composed of two parts: the ground model and the object model. In this model framework, the data from two sensor modalities play different roles. As shown in Figure 2b, the laser profile data are used as the main cues for building the ground model. However, the features of these depth profiles are too limited to identify object types. As shown in Figure 2c, edge profiles are extracted from the image and combined with associated laser profiles to form a multimodal representation of the object, which is used in building the object model. The ground model and the object model are unified in a Bayesian network, so that the optimal model parameters can be inferred to minimize uncertainties in this process. Figure 3 illustrates the multimodal profile model in the form of a Bayesian network. Using this model, the moving scene analysis becomes a problem of maximizing the joint probability distribution of all of the random variables involved. This joint probability distribution can be derived as shown in Equations (1)–(4).

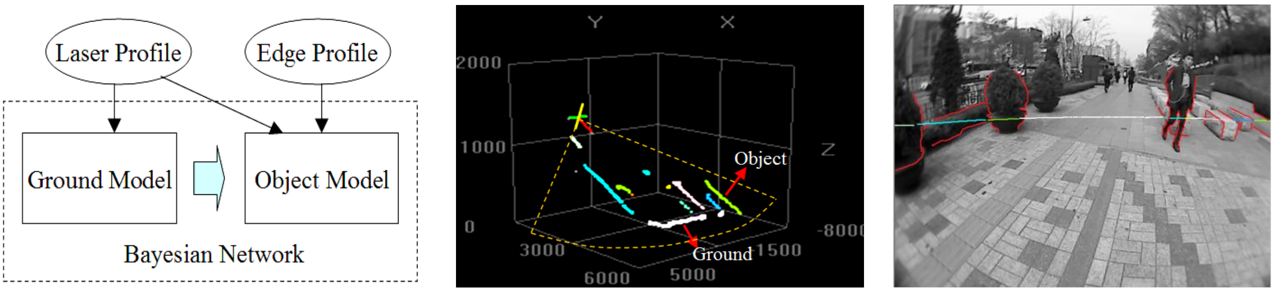
$$\mathrm{P}(G^t, G_e^t, Q^t, \mathit{\Psi}^t) = \mathrm{P}(G^t, G_e^t, Q^t)\,\mathrm{P}(\mathit{\Psi}^t \mid G^t, G_e^t, Q^t) \tag{1}$$

$$\begin{aligned} \mathrm{P}(G^t, G_e^t, Q^t) &= \mathrm{P}(G^t)\,\mathrm{P}(G_e^t \mid G^t)\,\mathrm{P}(Q^t \mid G^t, G_e^t) \\ &= \mathrm{P}(G^t)\,\mathrm{P}(G_e^t \mid G^t)\,\mathrm{P}(Q^t \mid G^t) \end{aligned} \tag{2}$$

$$\mathrm{P}(\mathit{\Psi}^t \mid G^t, G_e^t, Q^t) = \sum_{m=1}^{n} \mathrm{P}(\mathit{\Psi}_m^t \mid Q^t) \tag{3}$$

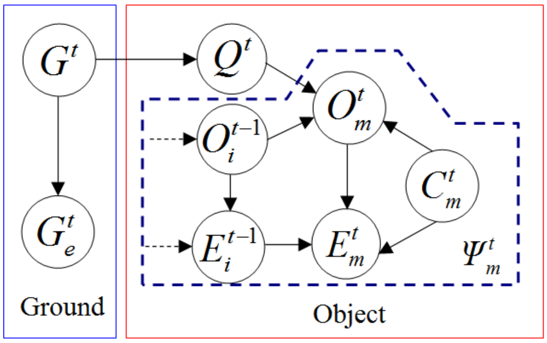
$$\begin{aligned} \mathrm{P}(\mathit{\Psi}_m^t \mid Q^t) = {} & \mathrm{P}(O_i^{t-1})\,\mathrm{P}(O_m^t \mid Q^t)\,\mathrm{P}(O_m^t \mid O_i^{t-1}) \\ & \cdot \mathrm{P}(E_i^{t-1} \mid O_i^{t-1})\,\mathrm{P}(E_m^t \mid O_m^t)\,\mathrm{P}(E_m^t \mid E_i^{t-1})\,\mathrm{P}(C_m^t \mid E_m^t, O_m^t) \end{aligned} \tag{4}$$

Figure 2. Overview of multimodal profile model. (**a**) General diagram; (**b**) laser range data; (**c**) fusion of the image and laser data.



(**a**) (**b**) (**c**)

Figure 3. Multimodal profile model.



Variable	Description
G^t	Ground at time *t*
G_e^t	Ground evidence from laser data at time *t*
Q^t	Assignment of laser data to object at time *t*
O_m^t	*m*-th object instance at time *t*
O_i^{t-1}	*i*-th object instance at time *t*-1
E_m^t	Edge distribution of *m*-th object at time *t*
E_i^{t-1}	Edge distribution of *i*-th object at time *t*-1
C_m^t	Assignment of *m*-th object to categories at *t*

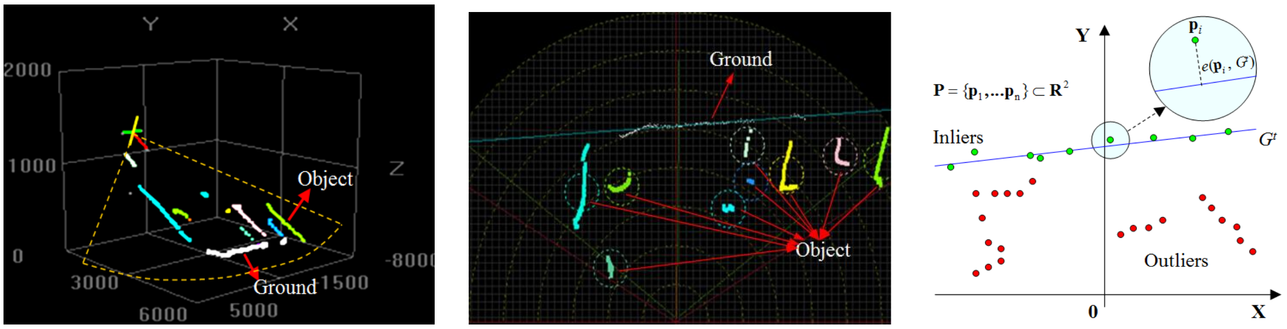
In Equation (1), $\mathrm{P}(G^t, G_e^t, Q^t)$ is the joint probability of ground location and laser points labeling, and $\mathrm{P}(\Psi^t \mid G^t, G_e^t, Q^t)$ denotes the joint probability of all objects Ψ^t given ground location and laser data labeling. In Equation (2), $\mathrm{P}(G^t)\mathrm{P}(G_e^t \mid G^t) \propto \mathrm{P}(G^t \mid G_e^t)$ is the posterior probability of ground G^t by observing ground evidence G_e^t, and $\mathrm{P}(Q^t \mid G^t, G_e^t)$ gives the probability of assigning laser data to the object given estimated ground G^t. Because Q^t and G_e^t are conditional independent given G^t, $\mathrm{P}(Q^t \mid G^t, G_e^t)$ is equivalent to $\mathrm{P}(Q^t \mid G^t)$. In Equation (3), the joint probability of all objects is approximated as the total probability of each object instance. Here, we assume that the probability of each individual object is independent.

In Equation (4), $\mathrm{P}(O_m^t \mid Q^t)$ is the probability of the *m*-th object instance given object laser point labeling Q^t, $\mathrm{P}(O_m^t \mid O_i^{t-1})$ is the temporal correlation of the *m*-th object instance, $\mathrm{P}(E_m^t \mid O_m^t)$ indicates the fusion of the *m*-th object's laser profile and the edge profile in frame *t*, $\mathrm{P}(E_m^t \mid E_i^{t-1})$ is the temporal correlation of the *m*-th object's edge distribution and $\mathrm{P}(C_m^t \mid E_m^t, O_m^t)$ indicates the object's type assignment probability given the *m*-th object's multimodal profile. In the following sections, each probability distribution in the model will be defined, and the methods used to find their maximum probability values will be discussed.

3.2. Ground Estimation

As shown in Figure 4a, because the laser scanner points downward at a fixed angle, some of the laser points are reflected from the ground, and the others are reflected from the objects. Figure 4b shows laser data observed in a 2D scanning frame. Assuming that the ground plane has a low curvature, a linear model is used to approximate the ground plane geometry. As shown in Figure 4c, in the laser's scanning frame, the ground at time *t* is defined as G^t: g_1x + g_2y + g_3 = 0. To find the G^t that is the most probable to be the real ground G^{t*}, a MAP (Maximum-a-Posteriori) estimation of G^{t*} is searched as $\arg\max_{G^t} \mathrm{P}(G^t \mid G_e^t)$. Using the Bayesian rule, the posterior probability $\mathrm{P}(G^t \mid G_e^t)$ can be derived as $\mathrm{P}(G^t \mid G_e^t) \propto \mathrm{P}(G^t)\mathrm{P}(G_e^t \mid G^t)$, where $\mathrm{P}(G^t)$ is the prior probability of G^t and $\mathrm{P}(G_e^t \mid G^t)$ is the likelihood of a given G^t by observing ground evidence G_e^t.

Figure 4. Ground estimation in laser scanning frame. (**a**) Laser data in 3D frame; (**b**) laser data in 2D scanning frame; (**c**) ground model definition.



(**a**) (**b**) (**c**)

Ground evidence G_e^t is measured through the point-wise error $e_i(\mathbf{p}_i, G^t)$, which is defined as the Euclidean distance between the laser point and the hypothesized line model. As illustrated in Figure 4c, if G^t is the ground, then the laser points on the ground can be regarded as the "inliers" of this G^t, and the laser points on objects can be regarded as the "outliers" of G^t. With this observation, the probability distribution of $e_i(\mathbf{p}_i, G^t)$ is modeled as a mixture of two parts, as in Equation (5). The inlier error is modeled using a Gaussian distribution, and the outlier error is modeled as a uniform distribution. The two parts are combined by using a radio parameter r. Assuming that $e_i(\mathbf{p}_i, G^t)$ at each point is independent from the others, $\mathrm{P}(G_e^t \mid G^t)$ can be defined as shown in Equation (6).

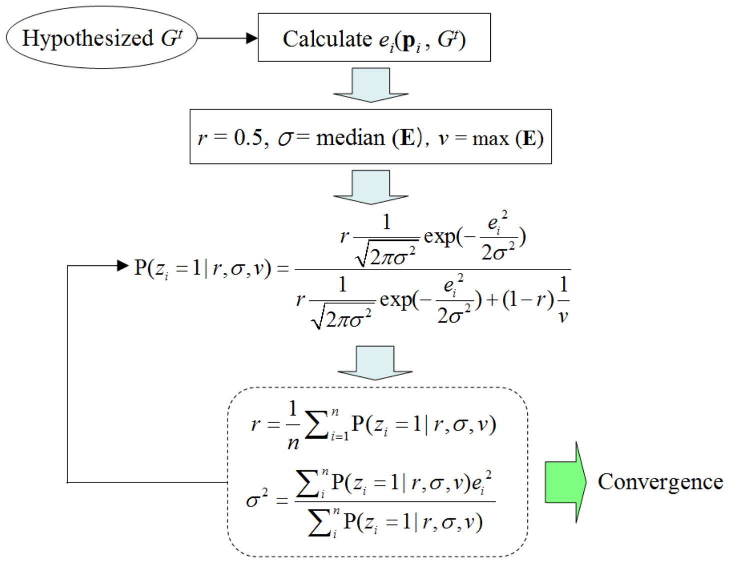
$$\mathrm{P}(e_i) = r\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{{e_i}^2}{2\sigma^2}) + (1-r)\frac{1}{v} \tag{5}$$

$$\mathrm{P}(G_e^t \mid G^t) = \prod_{i=1}^{n}\mathrm{P}(e_i) \tag{6}$$

To calculate $\mathrm{P}(e_i)$ using the mixture probability distribution defined in Equation (5), three parameters (r, σ, v) must be determined. Given a set of laser point data, the EM (Expectation-Maximization) algorithm is used here to fit (r, σ, v). As Figure 5 shows, for a given model hypothesis G^t, the set of all point errors $\mathbf{E} = \{e_1, e_2, \ldots e_n\}$ is used to initialize (r, σ, v), which is then updated iteratively by calculating the probability of each laser point belonging to the ground. Here, a set of indicator variables z_i is introduced; $z_i = 1$ indicates that the i-th data point is from the ground, and $z_i = 0$ means that it is from the object. Here, z_i is treated as missing data, and its probability values are updated iteratively together with (r, σ, v) in order to approach the best values. This iterative procedure is repeated until convergence.

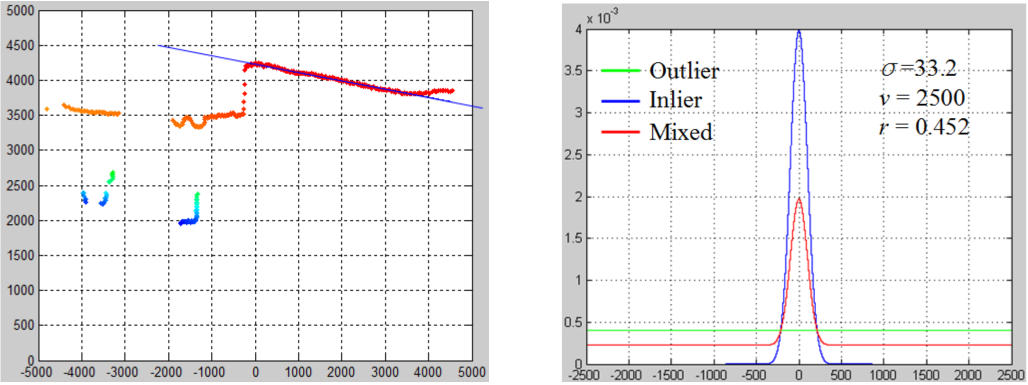
A random sampling scheme is used to form the ground model hypothesis. Two laser points are randomly selected in the laser data frame to give a hypothesized line model G^t. For each G^t, $\mathrm{P}(G_e^t \mid G^t)$ is calculated using $\mathrm{P}(e_i)$, the pdf (Probability Density Function) of which can be fitted through the above EM procedure. Finally, among all hypothesized G^t, the one with the maximum $\mathrm{P}(G_e^t \mid G^t)$ is chosen as the ML (Maximum Likelihood) estimation of the real ground.

Figure 5. EM procedure for fitting $P(e_i)$.


Hypothesized G^t
Calculate $e_i(\mathbf{p}_i, G^t)$
$r = 0.5$, σ = median (**E**), v = max (**E**)
$\mathrm{P}(z_i = 1 | r, \sigma, v) = \frac{r \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{e_i^2}{2\sigma^2})}{r \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{e_i^2}{2\sigma^2}) + (1-r)\frac{1}{v}}$
$r = \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}(z_i = 1 | r, \sigma, v)$
$\sigma^2 = \frac{\sum_i^n \mathrm{P}(z_i = 1 | r, \sigma, v) e_i^2}{\sum_i^n \mathrm{P}(z_i = 1 | r, \sigma, v)}$
Convergence

The results of ground estimation are shown in Figure 6. Figure 6a shows the best ground line model G^{t*} among all hypothesized G^t, and Figure 6b shows the fitted error probability distribution with respect to this G^{t*}. Moreover, when EM converges, a ground probability map can be obtained using $\mathrm{P}(z_i = 1)$, as is denoted by the heat color indicated in Figure 6a. This ground probability map is a valuable cue for labeling laser data points as in the ground or in objects.

Figure 6. An example of the ground estimation result. (**a**) Ground probability map; (**b**) fitted error distribution.



(**a**) (**b**)

To obtain a MAP estimation of the ground, instead of using random sampling, a guided sampling scheme is used to generate hypothesized G^t. Here, each data point is assigned a sampling weight to denote its importance in the sampling process. An initial sampling weight is assigned using a 2D Gaussian distribution, as in Equation (7), where (x, y) is a point in the laser's 2D scanning frame and (x_0, y_0) is the ground center position obtained from prior knowledge. For the frames following the initial frame, $\mathrm{P}(G^t)$ is modeled using the posterior probability of the ground model at time t−1. The final prior probability $\mathrm{P}(G^t)$ is defined as in Equation (8). In the initial frame, $\mathrm{P}(G^t)$ is determined by the initial sampling weight of two laser points. In the following frames, P(G^t) is determined by the posterior probability of the ground model in the previous frame.

$$\mathrm{P}(x,y) = A\exp(-(\frac{(x-x_0)^2}{2\sigma_x^2}+\frac{(y-y_0)^2}{2\sigma_y^2})) \tag{7}$$

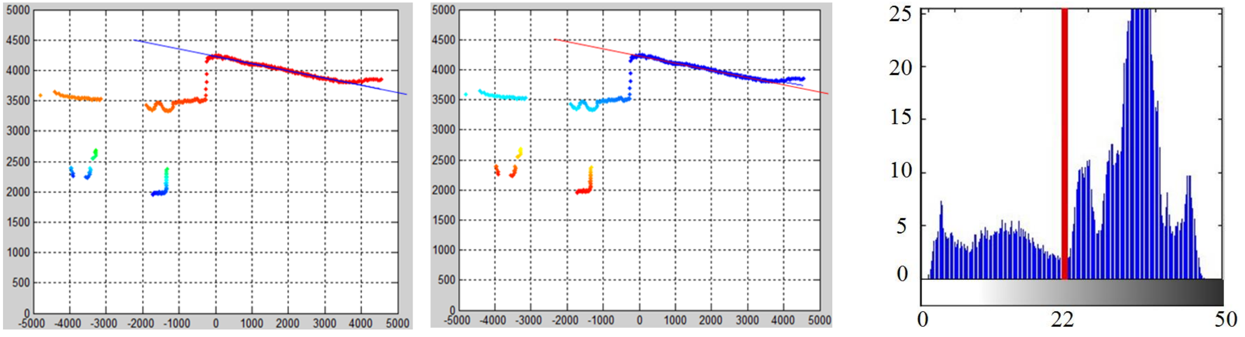
$$\mathrm{P}(G^t) = \begin{cases} \mathrm{P}(x_1,y_1) * P(x_2,y_2) & t=0 \\ \mathrm{P}(G^{t-1} \mid G_e^{t-1}) & t>0 \end{cases} \tag{8}$$

3.3. Object Detection

3.3.1. Object Probability Map Labeling

After the best ground model G^{t*} is found, P(z_i = 0) gives the probability of each laser point belonging to the object. By using P(z_i = 0), an object probability map can be obtained, which is a reversed version of the ground probability map, as shown in Figure 7b. Using this object probability map, the problem of object detection can be modeled as seeking the optimal labeling of laser points as belonging to objects or the ground.

Figure 7. Object probability map. (**a**) Ground probability map; (**b**) object probability map; (**c**) optimal labeling threshold.



(**a**) (**b**) (**c**)

In the multimodal profile model, Q^t is used to denote such a labeling. Searching for the optimal labeling emerges as another MAP estimation problem as: $Q^{t*} = \arg\max_{Q^t} \mathrm{P}(Q^t \mid G^t)$. By applying the Bayesian rule, the posterior probability $\mathrm{P}(Q^t \mid G^t)$ can be written as $\mathrm{P}(Q^t \mid G^t) \propto \mathrm{P}(Q^t)\mathrm{P}(G^t \mid Q^t)$. Assuming each labeling is equally likely to be Q^{t*}, P(Q^t) can be omitted. The P(G^t | Q^t) is measured using a between-class variance σ_{og}^2. To define this σ_{og}^2, a histogram of the object probability map is built as shown in Figure 7c. The horizontal axis is −log(P(z_i = 0)) ranging from zero to 50, and the vertical axis is the number of laser points that fall in each bin. Based on this minus-log-probability histogram, the hypothesized labeling Q^t is formed by choosing one of the bins as a threshold to divide the laser points into either the object or the ground class, and the between-class variance σ_{og}^2 can be calculated as in Equation (9), where r_{obj} and r_{gnd} are the ratio of the object laser points and the ground laser points and μ_{obj} and μ_{gnd} are the mean minus-log-probability of the object class and ground class. Finally, the Q^t with the largest σ_{og}^2 is selected as the optimal labeling Q^{t*}. In Figure 7c, the bin with the value “22” is selected as the best threshold. The minus-log-probability of 22 corresponds to a probability value of 0.82. Therefore, in the object probability map, a laser point with a probability value larger than or equal to 0.82 is labeled as belonging to the object.

$$\sigma_{\mathrm{og}}^2 = r_{\mathrm{obj}} r_{\mathrm{gnd}} (\mu_{\mathrm{obj}} - \mu_{\mathrm{gnd}})^2 \tag{9}$$

3.3.2. Laser Profile Clustering

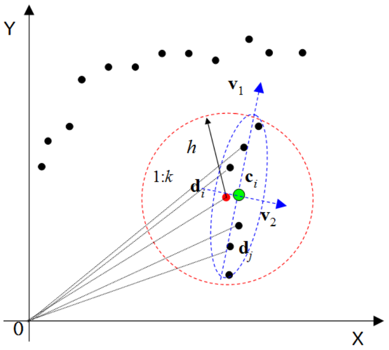
Object detection at a high resolution is a highly uncertain process. A MAP estimation of $\mathrm{P}(Q^t \mid G^t)$ can reduce uncertainties when separating object laser points from ground laser points. In order to handle uncertainties when separating individual objects from each other, the maximum value of $\mathrm{P}(O_m^t \mid Q^t)$ is to be searched. Assuming that the object generally has a smooth surface profile, then $\mathrm{P}(O_m^t \mid Q^t)$ can be measured according to the degree of smoothness observed in the laser data. This turns into a laser profile clustering problem. Here, we solve this problem based on a mean shift clustering scheme.

Mean shift clustering can be considered as a nonparametric kernel density estimation [29,30], where an unknown density function f(x) of the data is estimated via a kernel density function $\hat{\mathrm{f}}(x)$. The local maxima of the estimated kernel density function then yields cluster centers of the data. A general form of the kernel density function is given by Equation (10), where K is the kernel function and h is the kernel size.

$$\hat{\mathrm{f}}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{10}$$

Based on this kernel density estimation scheme, the kernel density function is defined to determine the smooth laser clusters. Now, given a set of object laser points $\mathbf{P} = \{\mathbf{p}_1, \ldots \mathbf{p}_n\}$, the goal is to cluster a smooth profile in this set of laser points. Here, a local smooth likelihood function $\mathrm{S}_i(\mathbf{x})$ is defined in a neighborhood of $\mathbf{p}_i$ to estimate the probability that a point $\mathbf{x}$ is located on a local object profile in this neighborhood. To define $\mathrm{S}_i(\mathbf{x})$, a combination of two kernels is used.

Figure 8. Definition of circular and ellipse kernels.



Variable	Description
$\mathbf{c}_i$	Ellipse kernel centroid
$\mathbf{p}_i$	Flat kernel centroid
h	Flat kernel bandwidth
$\mathbf{p}_j$	Laser points inside flat kernel
$\mathbf{v}_1$, $\mathbf{v}_2$	Ellipse kernel axis
k	Number of laser points inside kernel

As is shown in Figure 8, the flat circular kernel determines the scale of the neighborhood to be evaluated, depending on kernel size h. All laser points that are included in the circular kernel are used to calculate an ellipse kernel $\mathbf{E}_i$ by applying PCA (principal component analysis) to the involved laser points, as in Equation (11), where $\mathbf{c}_i$ is the centroid and $\boldsymbol{\Sigma}_i$ is a covariance matrix with two eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$. Because $\mathbf{v}_2$ is the direction in which the smallest depth variance occurs, its orthonormal vector $\mathbf{v}_1$ indicates the most probable position of a local object profile in this neighborhood.

$$\mathbf{E}_i(\mathbf{x}) = \{\mathbf{x} : (\mathbf{x}-\mathbf{c}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\mathbf{c}_i) \le 1\} \tag{11}$$

$$\mathrm{S}_i(\mathbf{x}) = \mathrm{N}_i(\mathbf{x},\mathbf{c}_i,\boldsymbol{\Sigma}_i)[h^2 - [(\mathbf{x}-\mathbf{c}_i)\cdot\mathbf{v}_{2i}]^2] \tag{12}$$

The smoothness likelihood $\mathrm{S}_i(\mathbf{x})$ of this local neighborhood is then evaluated with respect to $\mathbf{v}_1$ as in Equation (12). The squared distance from a position $\mathbf{x}$ to the possible object profile location $\mathbf{v}_1$ is used to measure the smoothness, which reflects the probability of a position $\mathbf{x} \in \mathbf{R}^2$ located on this profile. A 2D anisotropic Gaussian weighting function is used to diminish the possibility of distant points on the local profile. For each sampled laser point $\mathbf{p}_i$, a corresponding $S_i(\mathbf{x})$ can be calculated, which is illustrated in Figure 9. By accumulating all of the $S_i(\mathbf{x})$ defined at each laser point in the neighborhood, a complete kernel density function $\hat{\mathrm{f}}(x)$ can be obtained as in Equation (13), where K is a normalizing factor. This kernel density function $\hat{\mathrm{f}}(x)$ is used to approximate the unknown object density $\mathrm{P}(O_m^t \mid Q^t)$.

$$\mathrm{P}(O_m^t \mid Q^t) \propto \hat{\mathrm{f}}(\mathbf{x}) = \frac{1}{K}\sum_{i=1}^{n}\mathrm{S}_i(\mathbf{x}) \tag{13}$$

Figure 9. Local smooth likelihood function.

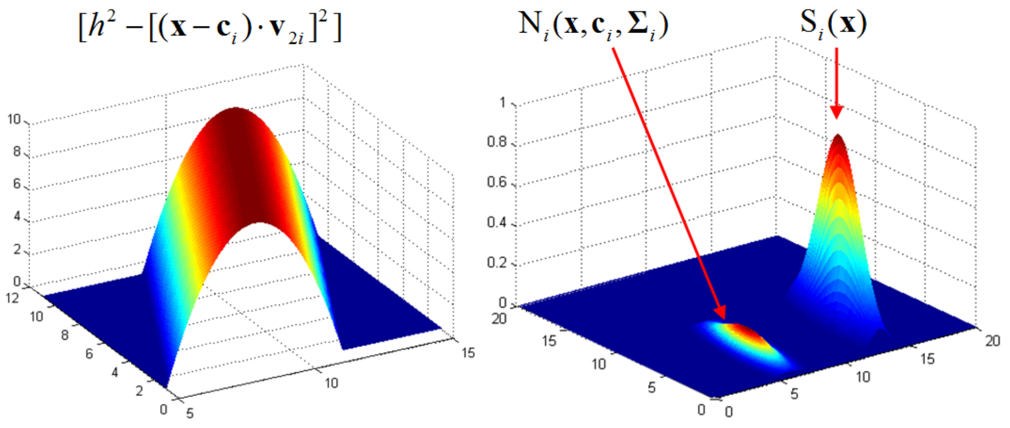
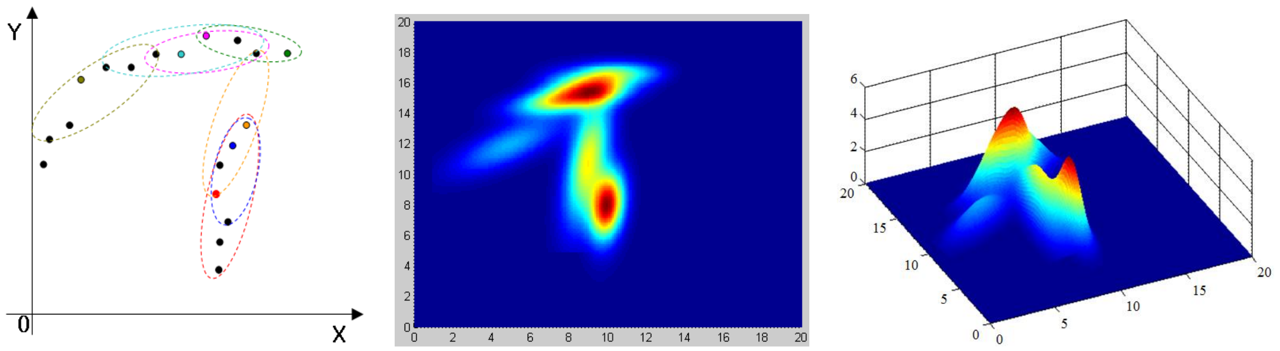


Figure 10. Accumulated smooth likelihood function. (**a**) Ellipse kernel in the laser frame; (**b**) accumulated smooth likelihood function in the 2D view and 3D view.



(**a**) (**b**) (**c**)

As shown in Figure 10, the local smoothness likelihood $S_i(\mathbf{x})$ defined at each ellipse kernel is accumulated to approximate the unknown pdf of $P(O_m^t \mid Q^t)$. Ellipse kernels that have similar positions and orientations will accumulate a higher probability vote on the positions involved, which reflects that they are more likely to locate on a smooth profile. The locations that get the largest vote are the cluster centers that indicate the existence of an object profile.

Using the defined $\hat{f}(x)$, the optimal segmentation of objects can be found by searching for local maxima in $\hat{f}(x)$. Here, a mean shift scheme based on gradient ascent maximization is used, as is defined in Equation (14). An iteration procedure from a sample point $\mathbf{p}_i$ converges if the mean shift vector is less than the given threshold.

$$\mathbf{p}_i^0 = \mathbf{p}_i, \quad \mathbf{p}_i^{k+1} = \mathbf{p}_i^k - \mathbf{m}_i^k$$
$$\mathbf{m}_i^k = \frac{\sum_{j=1}^{n} \mathrm{N}_j(\mathbf{p}_i^k - \mathbf{c}_j)\cdot[(\mathbf{p}_i^k - \mathbf{c}_j)\cdot \mathbf{n}_j]\cdot \mathbf{n}_j}{\sum_{j=1}^{n} \mathrm{N}_j(\mathbf{p}_i^k - \mathbf{c}_j)} \tag{14}$$

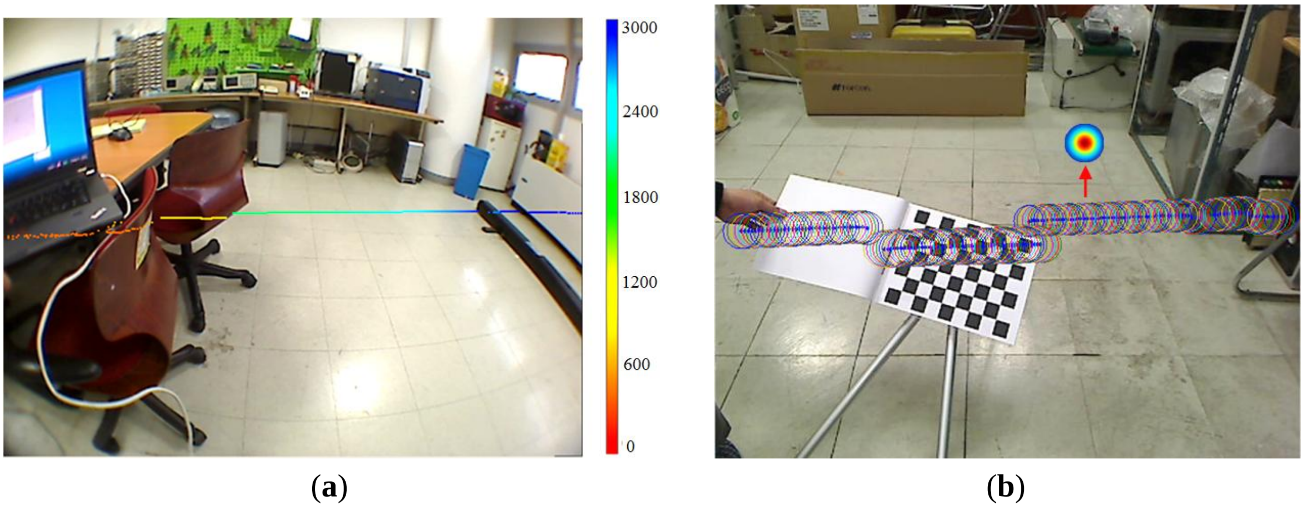
The above mean shift process is applied to every laser point in order to allow each one to converge to the nearest local maxima. After merging local maxima that are sufficiently close to one another, the final cluster centers can be determined. Each laser point is assigned to the cluster center to which it converges, and each laser profile cluster represents an individual object.

3.4. Multi-Category Object Classification

3.4.1. Pixel-Level Fusion of Laser and Image Data

Although laser data are efficient for ground and object detection, the depth profile they provide is still limited with regard to identifying object categories. Therefore, the edge profiles from images are used to supplement laser profiles in order to classify objects into multiple categories. To combine the edge profiles with the laser profiles, sensor calibration is performed to map laser profiles into the image domain.

Figure 11. Mapping laser points into the image frame. (**a**) Pixel-level fusion of laser and image data; (**b**) fusion uncertainties on the image plane.



(**a**) (**b**)

We use the method proposed in [31] to calibrate the laser sensor with the camera. After calibration, a mapping relationship is determined in order to map a laser point **P**$_l$ from the laser's coordinate system into the point **P**$_i$ in the image plane, as in Equation (15), where <**Φ Δ**> gives the mapping relationship between the laser and camera coordinates, **K** is the camera's intrinsic matrix and **d** is the non-linear distortion parameters. The example of a mapping result is shown in Figure 11a.

$$\mathbf{P}_{\mathrm{i}} = \mathrm{f}(\mathbf{P}_{\mathrm{l}}, \mathbf{\Phi}^{-1}, \mathbf{\Delta}, \mathbf{K}, \mathbf{d}) \tag{15}$$

In Equation (15), the set of calibration parameters <**Φ**, **Δ**, **K**, **d**> may contain errors. Therefore, uncertainties exist when using this parameter set to map **P**$_l$ to **P**$_i$. Here, we want to estimate the uncertainty of a non-linear mapping function $\mathrm{f}(\xi)$, based on the uncertainty of its mapping parameters $\xi = (\mathbf{\Phi}, \mathbf{\Delta}, \mathbf{K}, \mathbf{d})$. By using the non-linear covariance propagation theorem [32], the covariance matrix of $\mathrm{f}(\xi)$ can be approximated by Equation (16), where $\sum_{\xi}$ is the covariance matrix of calibration parameter vector ξ, and $\mathbf{J}_{\mathrm{f}}$ is the Jacobian matrix of $\mathrm{f}(\xi)$, evaluated at $\bar{\xi}$.

$$\sum\nolimits_{\mathrm{li}} = \mathbf{J}_{\mathrm{f}} \sum\nolimits_{\xi} \mathbf{J}_{\mathrm{f}}^{\mathrm{T}} \tag{16}$$

Given one laser point **P**$_{\mathbf{l}}$, a 2 × 16 Jacobian matrix **J**$_{\mathbf{f}}$ can be calculated with respect to the 16 calibration parameters. Thus, Equation (16) finally produces a 2 × 2 covariance matrix $\sum_{\mathrm{li}} = \mathrm{diag}(\sigma_{\mathrm{u}}^2, \sigma_{\mathrm{v}}^2)$, which gives the deviation of the coordinates of **P**$_i$ mapping from **P**$_l$. Figure 11b shows this pixel-level fusion uncertainty. The mapping position on the image is shown in blue dots, and the uncertainty of each mapped position is illustrated using an ellipse centered at the mapping position. The major and minor axes of the uncertainty ellipse are spanned by $3\sigma_{\mathrm{u}}$ and $3\sigma_{\mathrm{v}}$, representing 95% of the probability that the mapping position lies in this ellipse.

3.4.2. Profile-Level Fusion of Laser and Image Data

Based on the pixel-level fusion of the laser points and image pixels, we can further move to the profile-level fusion of the laser profile and edge profile. This multimodal profile fusion is represented as $\mathrm{P}(E_m^t \mid O_m^t)$ in the multimodal profile model.

Figure 12. Profile-level fusion of laser and image data.

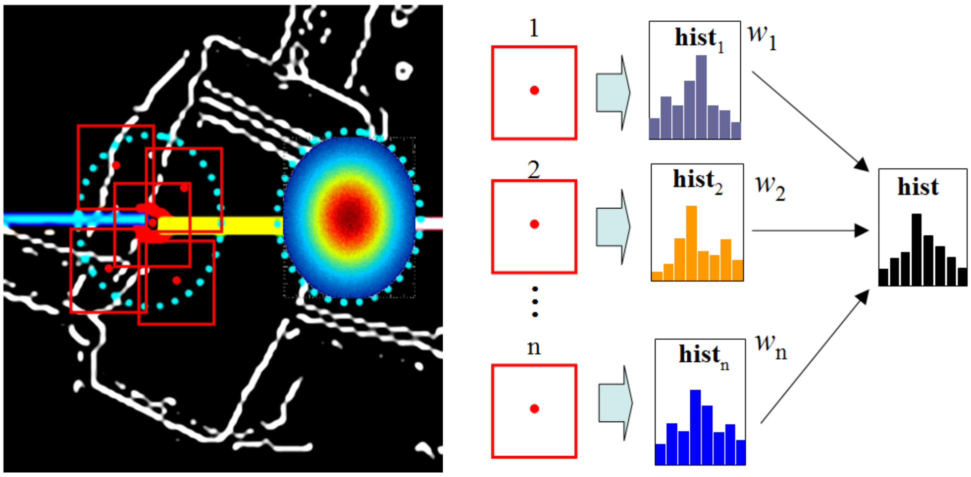


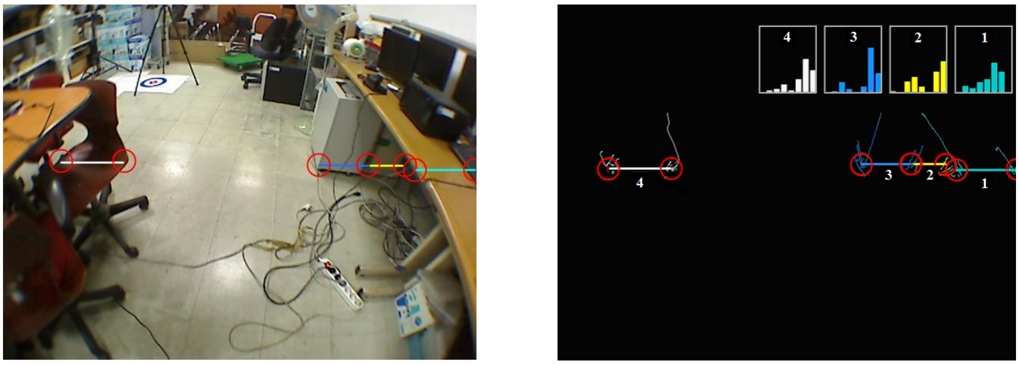
Figure 12 illustrates the modeling of $\mathrm{P}(E_m^t \mid I_m^t)$ in the form of a weighted edge-orientation histogram. In general, object laser profiles are supposed to break up at object boundaries. Therefore, if edge pixels are traced from these break-points on the image, useful local features concerning object shape can be obtained. In Figure 12, the image on the left side shows object laser profiles mapping onto the edge map, where $\sum_{\mathrm{li}}$ at each break point specifies an ellipse neighborhood, within which the mapping uncertainties can be modeled as a 2D Gaussian distribution.

As Figure 12 shows, inside this ellipse neighborhood, each sampled point specifies a rectangular window. For each rectangular window, a chain-code histogram is built to represent the edge orientation distribution around this neighborhood. Finally, all histograms are combined into one histogram, as in Equation (17), where $\mathbf{h}_i$ is a vector containing eight bins and w_i is a weight that is determined by the 2D Gaussian distribution that is used to model pixel-level fusion uncertainty.

$$\mathbf{H} = \sum\nolimits_{i=1}^{n} w_i \mathbf{h}_i \tag{17}$$

The weighted chain-code histogram can work effectively as a local shape descriptor. For objects with quasi-vertical side-profiles, a dual-peak pattern at Bin 2 and Bin 6 can be observed in the chain-code histogram, whereas for objects with lateral side-profiles, a dual-peak pattern at Bin 3 and Bin 7 can be observed. Therefore, the chain-code histogram can provide useful shape cues with which to identify vertical objects with lateral objects. An example of chain-code histogram extraction in a real scene is shown in Figure 13. The laser profile and its associated edge profile constitute a multimodal profile representation of the object.

Figure 13. An example of the weighted chain-code histogram.



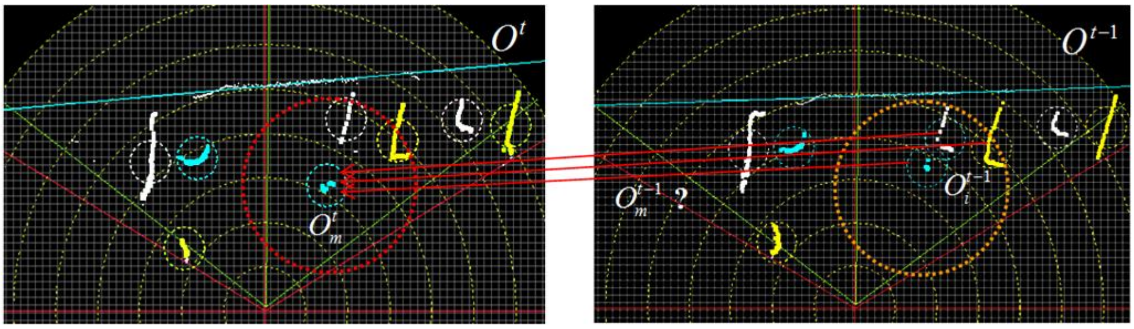
3.4.3. Temporal Correlation of Multimodal Profile Model

To further reduce uncertainties in moving scene analysis, temporal correlation cues across adjacent data frames are included in the multimodal profile model. The temporal correlation cues are represented as $\mathrm{P}(O_m^t \mid O_i^{t-1})$ and $\mathrm{P}(E_m^t \mid E_i^{t-1})$ in the multimodal profile model.

Here, a Bayesian tracker based on data association is proposed. As shown in Figure 14, there is a collection of *l* object profiles O^t at time *t* and a set of *k* object profiles O^{t-1} at time *t*−1. For one object profile O_m^t in O^t, decisions should be made regarding which object profile in O^{t-1} corresponds to O_m^{t-1} or whether no object profile corresponds to O_m^{t-1}, when O_m^t is a newly-emerging object. The posterior probability of each object profile in O^{t-1} being O_m^{t-1} is evaluated as in Equation (18), where $\mathrm{P}(O_i^{t-1})$ is the prior probability that each object profile in O^{t-1} is O_m^{t-1}, and $\mathrm{P}(O_i^{t-1} \mid O_m^t)$ is the likelihood of each object profile in O^{t-1}, given O_m^t as an on-line model.

$$\mathrm{P}(O_m^t \mid O_i^{t-1}) \propto \mathrm{P}(O_m^t)\mathrm{P}(O_i^{t-1} \mid O_m^t) \tag{18}$$

Figure 14. Laser profile association. (**a**) Laser profile in time *t*; (**b**) laser profile in time *t*−1.



(**a**) (**b**)

The prior probability term $\mathrm{P}(O_m^t)$ is modeled as a 2D isotropic Gaussian distribution $\mathrm{N}_m(\mathbf{x}, \mathbf{u}_m, \mathbf{\Sigma})$, where **x** is a position in the laser's scanning frame and **u**$_m$ is the centroid of laser profile O_m^t. The variance σ_x, σ_v is specified using the largest movement *T* of an object laser profile between adjacent frames. The likelihood of $\mathrm{P}(O_i^{t-1} \mid O_m^t)$ is calculated using the ICP (iterative closest point) algorithm. By applying ICP matching between model profile O_m^t and each target profile O_i^{t-1}, an optimal transformation $<\mathbf{R}, \mathbf{t}>$ can be obtained. In addition, a point-wise error can be calculated as $e_i = \left| \hat{\mathbf{x}}_i - \mathbf{R}\mathbf{x}_i - \mathbf{t} \right|$. Here, a threshold τ is applied on each e_j and the point $\mathbf{x}_j$, whose transformed error e_j is smaller than τ, is defined as an inlier, and the remaining points are all regarded as outliers. The number of inliers is used to define the likelihood, as in Equation (19), where N_{inlier} is the number of inliers in the target profile, S_{M} is the number of points in the model profile and S_{T} is the number of points in the target profile.

$$\mathrm{P}(O_i^{t-1} \mid O_m^t) = \frac{N_{\text{inlier}}}{\max(S_{\mathrm{M}}, S_{\mathrm{T}})} \tag{19}$$

By calculating the posterior probability of each target profile O_i^{t-1} as O_m^{t-1}, the target profile with the highest posterior is selected as O_m^{t-1}. Additionally, if none of the candidates has a posterior probability higher than the minimum threshold, then O_m^{t-1} is assumed to not exist, and O_m^t is deemed the newly emerging object profile. When $\mathrm{P}(O_m^t \mid O_i^{t-1})$ reaches the maximum by finding the O_m^{t-1}, $\mathrm{P}(E_m^t \mid E_i^{t-1})$ can also be regarded as reaching the maximum with $\mathrm{P}(E_m^t \mid E_m^{t-1})$, which is calculated as in Equation (20). An accumulated chain-code histogram can be built by propagating the chain-code histogram extracted around the corresponding laser profiles across frames.

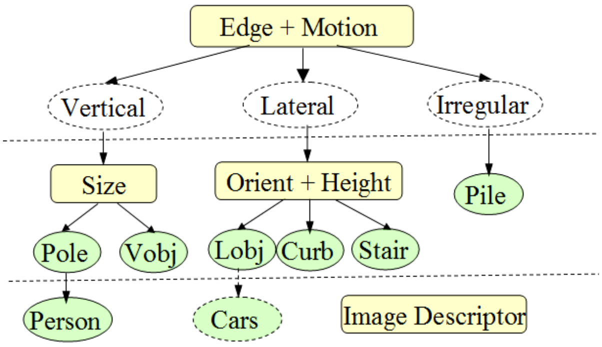
$$\mathrm{P}(E_m^t \mid E_m^{t-1}) \propto \eta\,\mathrm{P}(E_m^t) + (1-\eta)\,\mathrm{P}(E_m^{t-1}) \tag{20}$$

The accumulated chain-code histogram is more stable than the one built from the single frame is. When the laser scanner moves forward, the laser profile will project on different parts of one object. In real situations, some parts of this object might suffer from partial occlusion, whereas other parts might not. The accumulated chain-code histogram incorporates the local edge distribution from different local parts of the same object, thus greatly strengthening the histogram's tolerance to partial occlusion.

3.4.4. Object Classification

The chain-code histogram provides shape features from edge profiles. Size, height, orientation and vertical motion are extracted from laser profiles to constitute another four-bin histogram after normalization. Finally, a multimodal profile histogram can be obtained for object classification. A three-layer classification structure is proposed, as shown in Figure 15.

Figure 15. Three-layer structure for object classification.



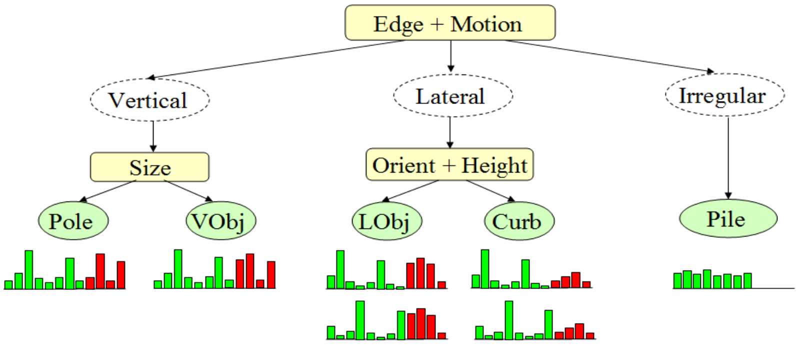
At the first level, three object types are defined. "Vertical Object (Vobj)" have quasi-vertical edges on their side-profiles, such as poles, pedestrians, doors and trees. "Lateral objects (Lobj)" have diagonally distributed edges on their side profiles, such as road curbs, side walls and guardrails. "Irregular objects" do not have regular shapes, such as bushes or piles of stones, and are addressed as "pile" in our system. Edge profiles are used as major features to classify objects in the first level, because it is more stable than the motion feature is, and it does not require the relative movement between the laser sensor and the object. However, if there is no sufficient edge profile that can be extracted around this object, then the motion feature from the laser profile is referred to in the classification process.

At the second level, objects are further classified into more specific categories by using features from the laser profile. "Poles" are vertical objects that have thin cross-sections, usually within a width of 30 cm. "Curb" usually has a lower height and quasi-horizontal profile orientation in the laser data domain. Stairs can also be identified as a specific object out of the "lateral object" group. After fitting the laser profile into a poly-line model, the number of poly-line segments is used as a criterion to identify stairs.

At the third level, a sophisticated image descriptor, such as the HOG (Histogram of Oriented Gradient) [33], is used to identify more specific object types from their "parent-categories". One of the benefits of this layered structure is that the recognition of more complex object types can be guided by the position of their parent-object instances. The location of their parent-object instance provides an ROI (Region of Interest) that can help reduce the recognition cost of complex object types. Here, in the current stage of our system, only pedestrian recognition is applied at the third level. A pre-trained HOG descriptor is used to match the ROI where the "pole" object is detected. If the matching score is higher than the pre-defined threshold, then the "person" is identified in the "pole" category.

In order to perform the classification using the above three-layered approach, the category prototypes of the defined object categories must be built from training data. In brief, after the collection of the training data concerning all object categories, the multimodal profile histogram is extracted for each object in the training data set, and then each multimodal profile histogram is treated as a 12-dimensional feature vector. K-means clustering is then performed in this 12-dimensional feature space to cluster the feature vectors into seven clusters. The cluster center is used as the object category prototype, as shown in Figure 16.

Figure 16. Object category prototype.



In the multimodal profile model, the probability of assigning an object type to a detected object is described by $\mathrm{P}(C_m^t \mid O_m^t, E_m^t)$. According to the Bayesian rule, this probability can be derived as in Equation (21), where $\mathrm{P}(O_m^t, E_m^t \mid C_m^t)$ is the likelihood of given object prototype *C*, and $\mathrm{P}(C_m^t)$ is a prior probability term. In general, it is assumed that all of the object categories are equally likely to appear on the road; therefore, $\mathrm{P}(C_m^t)$ can be omitted. The posterior probability is dependent on the likelihood term, $\mathrm{P}(O_m^t, E_m^t \mid C^t)$.

$$\mathrm{P}(C_m^t \mid O_m^t, E_m^t) = \frac{\mathrm{P}(O_m^t, E_m^t \mid C_m^t)\,\mathrm{P}(C_m^t)}{\mathrm{P}(O_m^t, E_m^t)} \tag{21}$$

$$\begin{aligned} &\mathrm{P}(C^t = i \mid \mathbf{H}_m^t) = \mathrm{Z}(\mathbf{H}_m^t)^{-1}\exp(-\mathrm{d}_i(\mathbf{H}_m^t)) \\ &\mathrm{Z}(\mathbf{H}_m^t) = \sum\nolimits_{i=1}^{K}\exp(-\mathrm{d}_i(\mathbf{H}_m^t)) \end{aligned} \tag{22}$$

The likelihood is modeled using the Gibbs distribution, as in Equation (22), where $\mathrm{d}_i(\mathbf{H}_m^t)$ represents the histogram distance from a measured multimodal profile histogram $\mathbf{H}_m^t$ to the *i*-th object category prototype histogram, and $\mathrm{Z}(\mathbf{H}_m^t)$ is a normalization factor. To calculate the histogram distance $\mathrm{d}_i(\mathbf{H}_m^t)$ between a measured histogram and a prototype histogram, the χ^2 distance measurement is used as in Equation (23), where $\mathbf{h}_0$ is the prototype histogram and $\mathbf{h}_m$ is the measured histogram. Parameter ***i*** represents the *i*-th histogram bin.

$$\chi^2(\mathbf{h}_m, \mathbf{h}_0) = \sum_i \frac{(\mathbf{h}_m[i] - \mathbf{h}_0[i])^2}{\mathbf{h}_m[i] + \mathbf{h}_0[i]} \tag{23}$$

4. Audio Message Generation

4.1. Walking Context Estimation

The walking context is defined as the safety level of the walking condition. A fuzzy logic inference model is built to evaluate this safety level, based on several factors defined in the output of the moving scene analysis, as illustrated in Figure 17. To define the membership functions of four input variables, sample data of these four variables are collected from various scenes. By examining the sampled data distribution in their respective domains, data clusters are localized and used to define fuzzy sets, as shown in Figure 18.

Figure 17. Fuzzy logic model for walking context estimation.

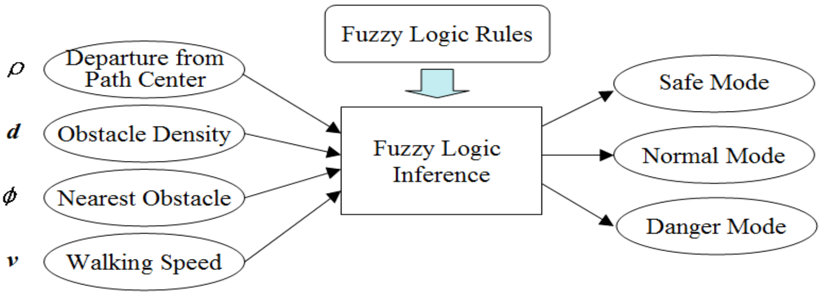
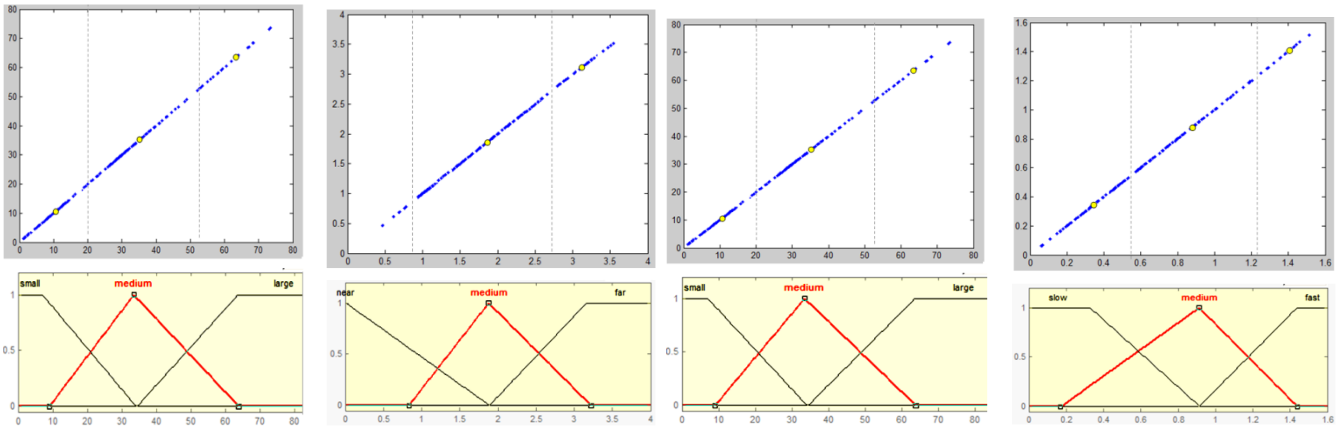


Figure 18. Fuzzification of walking context factors. (**a**) Fuzzification of ñ; (**b**) fuzzification of *d*; (**c**) fuzzification of ö; (**d**) fuzzification of *v*.



(**a**) (**b**) (**c**) (**d**)

The output variable of the walking context is defined as the degree of safety. It is denoted as ***c*** and measured as a percentage value. Because no direct data is available to calculate this value, the intuition and experience of a normal person is used to define its membership function. For each sampled scenario with a set of fuzzy values of four input variables, we let the users decide the degree of safety in this scenario by using their intuition and experience. With a set of fuzzy values of four input variables on one side and the user's answers on the safety degree on the other side, the parameters of a piece-wise linear function are learned from these data.

To perform fuzzy inference, a set of fuzzy rules is also needed. Basically, fuzzy rules take the form of "If...then..." in correlating a set of fuzzy values of input variables to a unique fuzzy value of the output variable. For example, "If ρ is low, and d is far, and ϕ is small and v is low, then the walking context c is safe". The fuzzy rules used in our fuzzy logic model are derived from the user-labeled data when the membership function of output variable c is learned. Therefore, these fuzzy rules reflect normal human decisions in a walking context under different situations. A list of the fuzzy rules is shown in Table S1, and a fuzzy inference process is illustrated in Figure S1. The result of all fuzzy inference processes can be represented as a solution space shown in Figure S2.

4.2. Message Definition

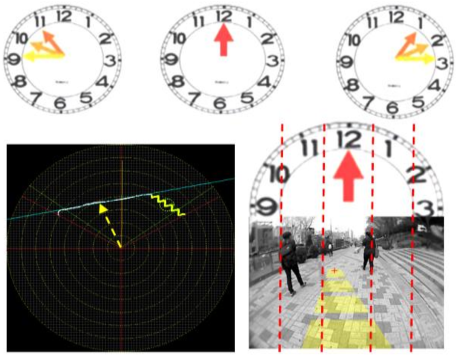
The message structure is defined based on the type of information it expresses. The information that should be delivered to the user is divided into four basic sets, as shown in Table 1. The words in brackets are template words that can be changed according to the result of the detection.

Table 1. Message definition.

Message Type	Message Example
Safe direction	"Go (12 o'clock)".
Walking context	"Danger context attention". "Back to normal/safe".
Closest object type	"vertical object", "lateral object", "pole", "person", "curb", "pile"
Closest object location	Acoustic beeper message
User motion	"Please walk faster/slow down/stop". "Large departure attention".

In the safe-direction message set, clock-face directions are used to give messages about direction, as shown in Figure 19. The laser's scanning plane is shown as a big clock-face, in which two leftward directions are specified as 10 and 11 o'clock and two rightward directions are at one and two o'clock. Twelve o'clock is used to represent a straight, forward direction. Object-type messages are defined based on the object classes described in Section 3.4.4. When many objects appear at the same time, the type of every object cannot be announced one by one. Considering that the object closest to the user is the most critical, only the closest object's type is delivered to the user. However, the location of the closest object is difficult to deliver using verbal messages, because it changes from frame to frame. Therefore, an acoustic beeper message set is defined to announce the nearest object position, which is similar to that in [16]. The definition of beeper messages are shown in Figure S3 and Table S2.

Figure 19. Definition of clock-face direction.



4.3. Message Delivery Scheme

To define a proper message delivery scheme, the estimated walking context offers valuable cues for selecting the messages that are more important at the moment. The set of message delivery rules is defined in Table 2.

Table 2. Message delivery rules.

Walking Context	Output Message Set
Safe	safe direction, user motion close obstacle type and position
Normal	safe direction, user motion, closest obstacle type and position
Danger	safe direction, walking context

In a "danger" context, it is urgent for the user to receive instantaneous messages about safe walking directions. Less important messages should be blocked in order to avoid their interference with the delivery of messages that give safe directions. In normal, safe contexts, it is more desirable for the user to know the types and positions of objects in the surrounding environment in order to maintain a safe walking direction. Therefore, in safe and normal contexts, a full set of message types can be delivered, whereas in dangerous contexts, only safe-direction and walking context messages are allowed to be delivered.

Another important aspect is the delivery timing. Here, the messages are divided into three types of timing, as shown in Table 3. Hard timing messages have a higher priority than soft timing messages do, and they must be delivered instantly whenever changes in the situation are detected. In hard timing messages, safe direction messages have a higher priority than walking context messages do. User motion messages have the lowest priority in these hard timing message sets.

Table 3. Message delivery timing.

Types of Timing	Instruction Set
Hard timing	safe direction > walking context > user motion
Soft timing	closest object type
Real timing	closest object position

Messages about the closest object's type are defined as soft timing messages. This type of message is considered less critical for the user's walking safety. Therefore, soft timing messages can be blocked by any hard timing message. Messages about the location of closest object are defined as real timing messages. Because an acoustic beeper is used to indicate object location in real time, this type of beeper message can be delivered in parallel with other verbal messages.

5. Experimental Results

To validate the effectiveness of the proposed system, it was tested using real-scene data collected from a laser scanner and a camera. In this section, the experimental results of each part of the system will be shown and discussed.

5.1. System Prototype and Test Data Collection

A prototype of the system is shown in Figure 20. A camera and a laser scanner were bound to each other and fixed on a solid shelf. The combined sensors were connected to a tablet computer for the acquisition and processing of the sensor data. The specifications of the sensors used in the experiment are shown in Table 4. A wearable design of the system prototype is shown Figure S6. In the experiment, the prototype system was mounted on a human who traversed various urban paths. Image frames with synchronized laser scan data on these urban path scenes were collected and recorded. Some test data samples are shown in Figure 21, and the system output on processing these test data is shown in Figure S4 and Figure S5.

Figure 20. System prototype.

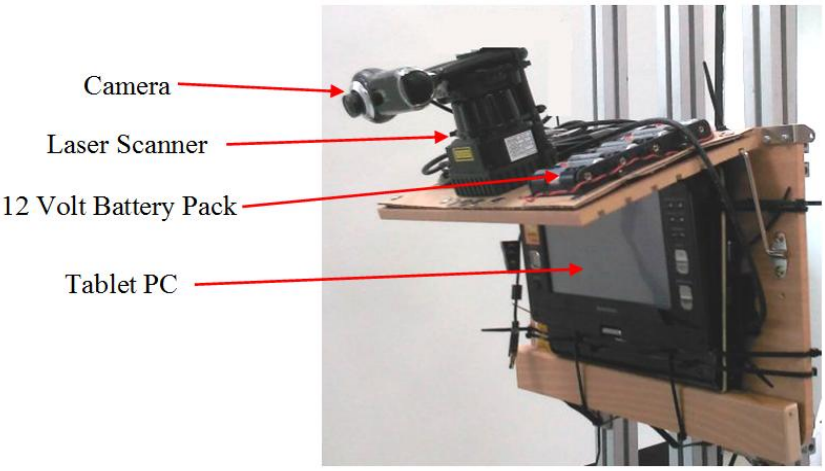


Table 4. Sensor specification.

System Component	Specification	
Hokuyo Laser Scanner UTM-30LX-EW	Angular resolution	0.25 °
	Range resolution	1 mm
	Scanning speed	25 ms/scan
	Scanning range	270 °
	Measuring range	0.1 m~30 m
Logitech Webcam 9000L	Image resolution	640 × 480
	Frame rate	30 fps

Figure 21. Test data samples.

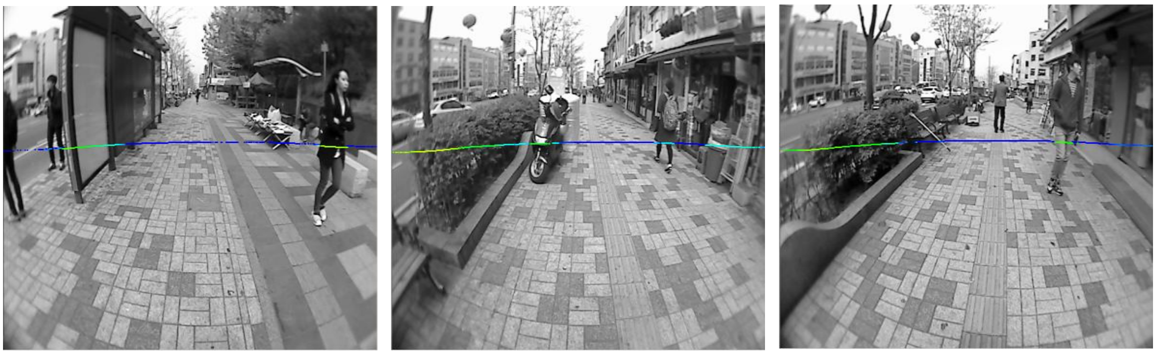
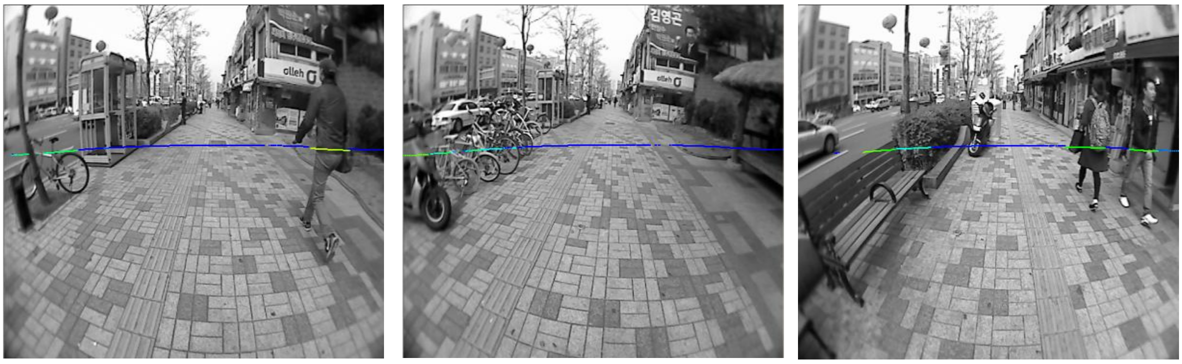


Figure 21. *Cont.*



5.2. Ground Estimation Performance

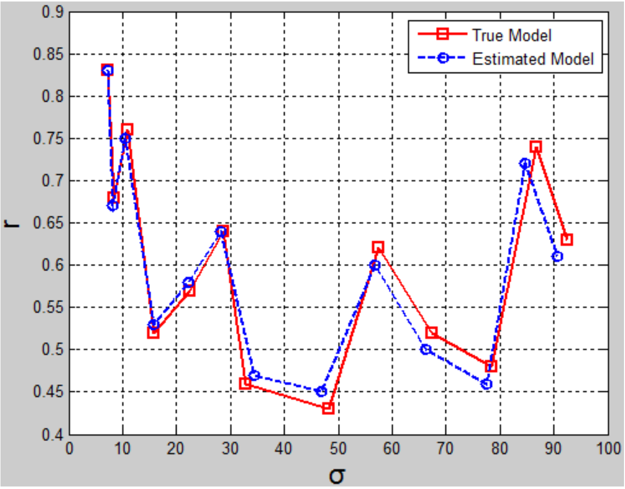
In this section, the ground estimation performance will be evaluated using the collected test data. Ground estimation is the most essential part of the whole system. It not only gives the ground location, but also provides a reference for generating hypotheses about the objects.

The performance of the ground estimation is analyzed according to two aspects: ground evidence model fitting and ground laser point detection rate. The ground evidence model refers to the Gaussian-uniform mixture type of error model. To evaluate its fitting performance, 13 representative scenarios were selected from test data. These 13 scenarios contained various ground conditions, from flat to rough, and various object densities, from low to high. The mixture type of error model was used to fit the laser scan data of these 13 scenarios. In the first round, the mixture model parameters were fitted from data using the EM algorithm. In the second round, the mixture model parameters were manually tuned to reflect real data distributions. The parameters of the manually tuned mixture model were used as true model parameters in order to evaluate how well model parameters were fitted.

In Figure 22, the best fitting parameters (r, *ó*) for each scenario in the first round are referred to as the “estimated model” marked with small circles. The parameters of the manually tuned model are referred to as the “true model” marked in small rectangles. Variance *ó* reflects the spreading range of the ground laser points, and the inlier ratio r reflects observable ground areas in the scene. It can be observed that in scenarios with small *ó* and large r, the estimated model parameters approached true model parameters very closely, whereas in scenarios with large *ó* and small r, more obvious deviation can be noticed between the estimated model and the true model.

From the results shown in Figure 22, it can be concluded that the Gaussian-uniform error model reflected the real data distribution well in cases of relatively flat ground with few objects. In cases of rough ground with many objects, the estimated *ó* tended to be smaller than the true *ó* value, which indicates a higher probability of identifying ground laser points as object laser points. This might give rise to the false negative rate in terms of ground laser point detection. Despite some increased inaccuracies in cases of rough ground, the errors were still within the acceptable range.

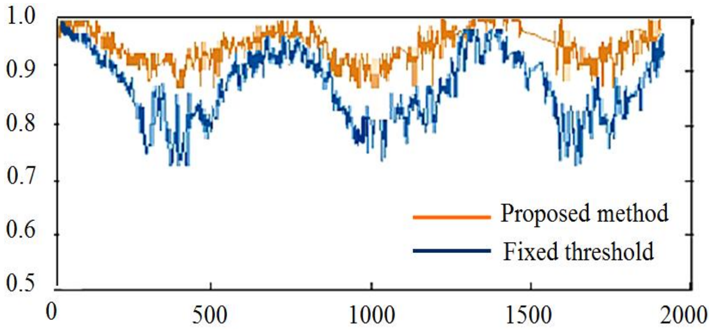
Figure 22. Ground evidence model fitting performance.



The ground point detection rate was evaluated using captured video sequence data. The results of detection from a representative video sequence that contains about 1876 data frames are shown in Figure 23. The horizontal axis represents consecutive image frames indexed from zero to 2000, and the vertical axis indicates the true positive rate, which is calculated as in Equation (24), where *TP* is true positives and *FN* is false negatives.

$$TPR = TP/(TP + FN) \tag{24}$$

Figure 23. Ground point detection rate.



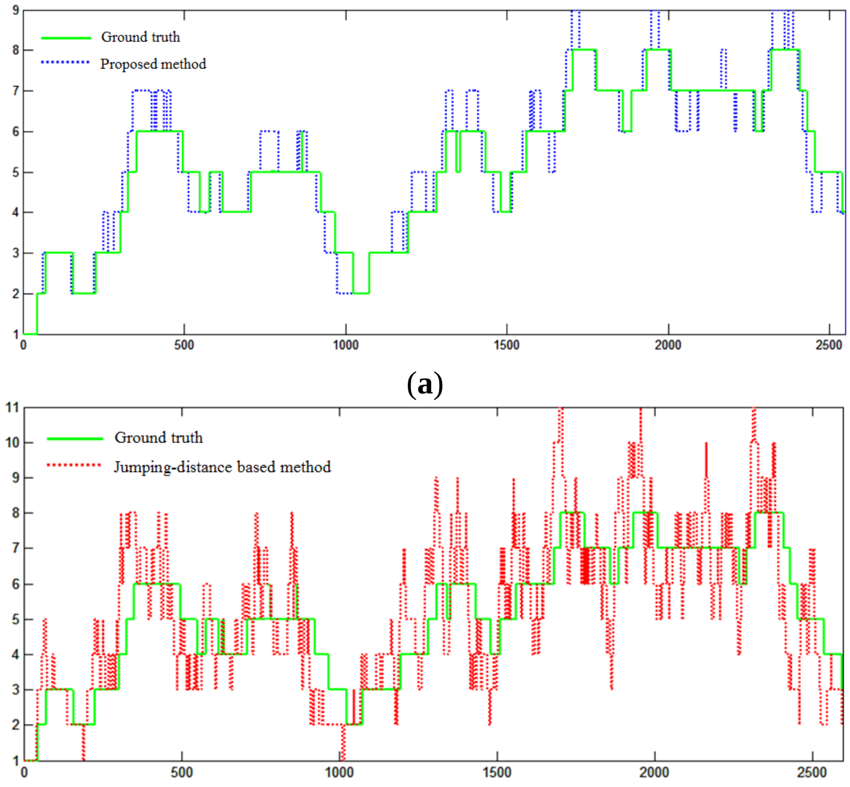
To calculate true positive and false negatives, the laser points of each data frame were manually labeled as ground points or object points. These manually labeled data are used as the ground truth. The results of detection in each data frame were compared with the ground truth; true positives were correctly detected ground points, and false negatives were ground points that were wrongly classified as object points. Of the two types of detection errors (false positive and false negative), here, we are more concerned with false positives. The proposed method tends to produce more false positive errors than false negative errors, which means that there is a higher chance for classifying ground points as object points than *vice versa*. This error producing property is also reasonable for reducing blind user's collision risk.

For a comparison, a method using a fixed threshold to determine ground points and object points was implemented. This fixed threshold method used a fixed distance threshold to the linear ground model to determine whether a laser point was a ground point or not. The true positive rate of ground points detected using the above two methods was calculated for every frame, and the results are shown in Figure 23. The results showed that the proposed method maintained a high detection rate through all frames compared with the fixed threshold method. Although the fixed threshold was tuned to achieve a high detection rate in the first few frames, its detection rate dropped dramatically, then increased and decreased vigorously in a very unstable manner. This result showed that the fixed threshold method could not adapt to accommodate various ground conditions. In contrast, the performance of the proposed method was stable in keeping a high detection rate across all frames.

5.3. Object Detection and Classification Performance

To evaluate the performance of object detection, the number of objects in the test video sequence was used as the major measurement. The results of detection from a representative video sequence containing 2670 data frames are shown in Figure 24a. The horizontal axis shows frame indexes, and the vertical axis shows the number of objects.

Figure 24. Object detection performance. (**a**) Proposed method performance; (**b**) jumping-distance method performance.



(**a**)

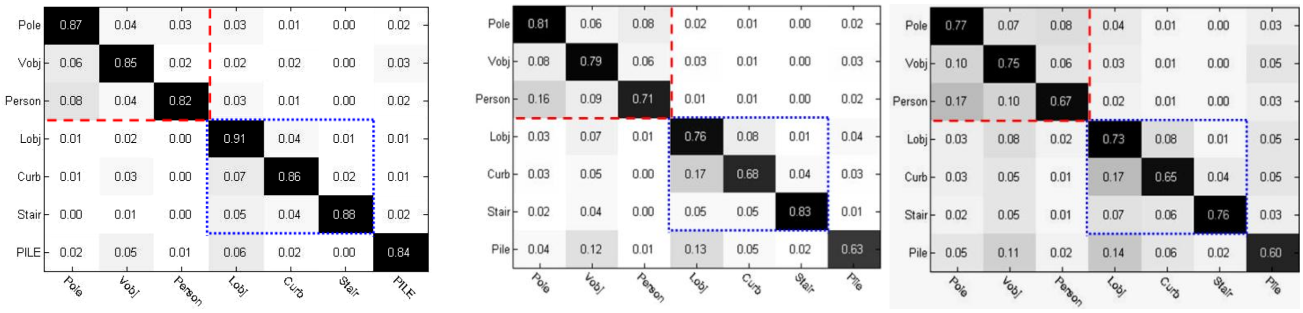
(**b**)

The ground truth data were obtained by manually labeling the number of objects in each data frame, and the number of objects detected in each frame was obtained by using the proposed clustering method. As Figure 24a shows, the proposed method tracked the ground truth data well in over 83% of the data frames, which indicates that the proposed smoothness likelihood function was able to capture smooth discontinuities by estimating laser data distributions in a local region.

For a comparison, a jumping distance-based method was implemented. The results of its object detection in the same video sequence are shown in Figure 24b. The jumping distance method used a fixed threshold to segment the laser profiles. As Figure 24b shows, the object detection accuracy of the jumping distance method was much worse than that of proposed method. There are not only large deviations from the ground truth, but also a large variance across frames. The result showed that compared with the proposed method, the jumping distance method tends to produce less accurate and unstable detection results.

The object classification performance was evaluated by using the confusion matrix shown in Figure 25. In the confusion matrix, the vertical axis shows the actual category and the horizontal axis shows the predicted category. The object types in each frame of the test video sequence were manually labeled as ground truth data. The results of classification from a representative test video sequence consisting of 2680 frames are shown in Figure 25.

Figure 25. Object classification performance. (**a**) Performance with proposed model; (**b**) Performance without temporal correlation; (**c**) Performance without temporal and spatial correlations.



	Pole	Vobj	Person	Lobj	Curb	Stair	PILE
Pole	0.87	0.04	0.03	0.03	0.01	0.00	0.02
Vobj	0.06	0.85	0.02	0.02	0.02	0.00	0.03
Person	0.08	0.04	0.82	0.03	0.01	0.00	0.02
Lobj	0.01	0.02	0.00	0.91	0.04	0.01	0.01
Curb	0.01	0.03	0.00	0.07	0.86	0.02	0.01
Stair	0.00	0.01	0.00	0.05	0.04	0.88	0.02
PILE	0.02	0.05	0.01	0.06	0.02	0.00	0.84

(**a**)

	Pole	Vobj	Person	Lobj	Curb	Stair	Pile
Pole	0.81	0.06	0.08	0.02	0.01	0.00	0.02
Vobj	0.08	0.79	0.06	0.03	0.01	0.00	0.03
Person	0.16	0.09	0.71	0.01	0.01	0.00	0.02
Lobj	0.03	0.07	0.01	0.76	0.08	0.01	0.04
Curb	0.03	0.05	0.00	0.17	0.68	0.04	0.03
Stair	0.02	0.04	0.00	0.05	0.05	0.83	0.01
Pile	0.04	0.12	0.01	0.13	0.05	0.02	0.63

(**b**)

	Pole	Vobj	Person	Lobj	Curb	Stair	Pile
Pole	0.77	0.07	0.08	0.04	0.01	0.00	0.03
Vobj	0.10	0.75	0.06	0.03	0.01	0.00	0.05
Person	0.17	0.10	0.67	0.02	0.01	0.00	0.03
Lobj	0.03	0.08	0.02	0.73	0.08	0.01	0.05
Curb	0.03	0.05	0.01	0.17	0.65	0.04	0.05
Stair	0.02	0.05	0.01	0.07	0.06	0.76	0.03
Pile	0.05	0.11	0.02	0.14	0.06	0.02	0.60

(**c**)

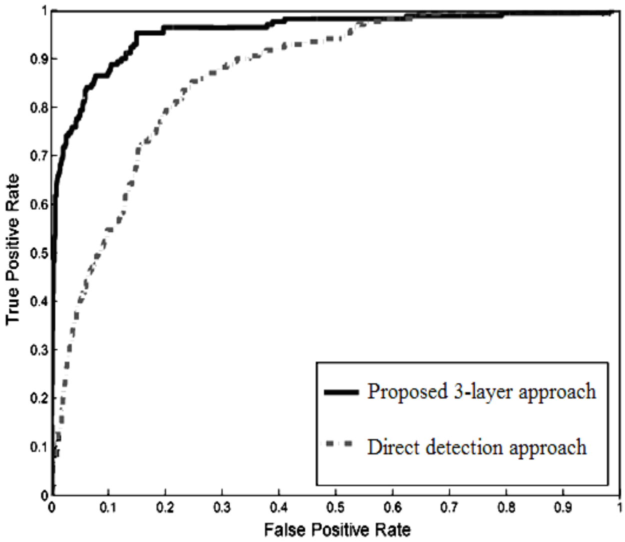
As shown in Figure 25a, the predicted object category labels were produced by using the full multimodal profile model, following the three-layer classification structure. As Figure 25a shows, the classification errors mainly occurred among the object types that belong to the same mid-level category, which is illustrated by the two sub-regions marked in Figure 25a. On the other hand, the confusion rate between the vertical object category and the lateral object category was relatively low. The weighted chain code histogram with temporal accumulation was effective in maintaining a low rate of classification errors in the first layer. This result provided a good basis for identifying finer object classes in the following layers. In fact, the major factor affecting the classification of mid-level categories in the first layer is the partial occlusion and overlapping of object edges in a local region. In the multimodal profile model, we proposed building edge histograms by fusing the edge distributions in multiple local regions both spatially and temporarily to reduce the influence of edge occlusion and overlapping.

To show the effectiveness of this multiple local region scheme, Figure 25b provides the results of classification obtained by using the incomplete multimodal profile model, without the temporal accumulation of the weighted chain-code histogram. Compared with the classification rate shown in Figure 25a, the results showed that the true positive rates in the diagonal dropped, and first-layer classification errors increased, which led to significant error propagation in the second and the third layers of object classes. A chain reaction initiated by the increase in classification errors among the first-layer categories resulted in the increase in total errors, which affected almost all classes.

In addition, Figure 25c shows the classification results using the direct fusion of edge and laser profiles. In this direct fusion method, the edge histogram was built using only a single, local region centered at the break point of the laser profile, without considering the edge distributions of other adjacent local regions or adjacent temporal frames. Built in this way, the edge histogram is much more easily affected when edge profile occlusion or overlapping occurs. The deterioration in the rate of object classification is clearly shown in Figure 25c. The chain reaction of severe error propagation caused a dramatic decrease in the true positive rates, with classification errors spreading to the false positive and false negative elements.

In the proposed three-layer classification structure, more specific object types with complex appearances can be identified using sophisticated image descriptors in the third layer. In the three-layer structure, the object class in the bottom layer benefited from its "parent category" in the upper layer. An example is pedestrian detection: if we use the HOG descriptor to detect pedestrians in the entire image, there will be many false positives, and the detection speed will be slow. However, if we use the multimodal profile model with a three-layer classification structure, the recognition of vertical objects will give reasonable constraints on the possible location of pedestrians. The HOG descriptor can be used to check only several vertical object locations for the recognition of pedestrians. This not only helps reduce false positives, but also hastens the detection of pedestrians.

Figure 26. Pedestrian detection performance.



The receiver operating characteristic (ROC) curve of the pedestrian detection rate is shown in Figure 26. The ROC was obtained by using the HOG descriptor with the three-layer structure in the multimodal profile model and by using the HOG descriptor alone to scan the entire image. The results clearly showed that pedestrian detection using the proposed three-layer structure outperforms the direct detection approach.

5.4. System Run Time Performance

The entire algorithm was implemented in C++ and tested on a laptop computer. The specifications of the testing platform are listed in Table 5. In Table 6, average runtime is shown with respect to each module, each function in the module, as well as each iteration unit of the function. The iteration unit of each function is shown in Table 7. The runtime of each function is calculated as the average runtime value on test data sets for one frame.

Table 5. Testing platform specification.

Platform Component	Specifications
CPU	Intel i5@ 2.5 GHZ
Memory	4GB DDR3
GPU	GeForce GT 740M
OS	Windows 7 32 bit
Programming Tool	Visual Studio 2010
Compiler	Microsoft VC++ 10

Table 6. Average run time performance.

Module	Function	Runtime/ Iteration Unit	Runtime/ Function	Runtime/ Module
Ground estimation	Ground model fitting	1.03 ms	7.34 ms	7.34 ms
Object detection	Laser profile clustering	0.12 ms	6.52 ms	21.25 ms
	Laser profile tracking	0.83 ms	14.73 ms	
Object classification	Multimodal profile histogram	0.62 ms	8.56 ms	38.83 ms
	Generic object classification	0.58 ms	6.82 ms	
	Pedestrian classification	4.87 ms	23.45 ms	
User feedback	Context estimation	5.35 ms	5.35 ms	8.75 ms
	Message generation	0.86 ms	3.4 ms	
Total	All functions		76.17 ms	76.17 ms

Table 7. Iteration unit of each function.

Function	Iteration Unit
Ground model fitting	One hypothesized model fitting
Laser profile clustering	One point converges to cluster center
Laser profile tracking	Tracking of one laser profile
Multimodal profile histogram	Get multimodal histogram from one object
Generic object classification	Classification of one generic object
Pedestrian classification	Classification of one pedestrian
Context estimation	Estimate context for one frame
Message generation	Generate one message

The indicator “Runtime per iteration unit” shows that it takes about 1.2 ms (0.62 ms + 0.58 ms) to recognize one generic object type. Generic object types are those in the first and second layer of the three-layer structure, as shown in Figure 15. For pedestrian detection in the third layer, it takes 4.87 ms to calculate the HOG descriptor and to do the classification for one hypothesized window. When calculating the HOG descriptor, a GPU accelerated version was used, which is much faster than a CPU-based version. The total average run-time was 76.17 ms for one data frame, which is approximately 13 fps.

In our experiment, a blind pedestrian was observed to walk at a speed of around 0.5 m/s~1.8 m/s on average. Under this walking speed, 3~5 s was an appropriate time interval for message delivery. Therefore, to meet the run-time requirement of the guidance task, a minimum processing speed of 5 fps was required. With non-optimized code in debugging mode, the proposed system can run at approximately 10~15 fps. This shows that the proposed system could fully satisfy the real-time requirements for general guidance tasks. For the maximum traveling speed at which the runtime performance of the system is still reasonable, it actually depends on the time it takes to read out one verbal message to the user; because the time it takes to read out one verbal message to the user is much longer than the time it takes for data acquisition and processing. For example, it normally takes 1 s to finish reading one verbal message defined in Table 1. Given that data acquisition and processing costs an additional 0.1 s, then the system response time will add up to 1.1 s. Assuming sensors can detect objects at 5 m ahead of the user, the maximum traveling speed of the user can be estimated as 4.54 m/s. This maximum speed is already over two-times faster than normal pedestrian speed.

6. Conclusions

In this paper, an audio guidance system based on context awareness is proposed to assist blind people in traversing local pathways safely and efficiently. The proposed multimodal profile model greatly improved the resolution of scene interpretation by enabling multi-category object identification. An audio message interface was proposed to deliver highly semantic information to the user, based on a walking context. The proposed system has the following limitations. In ground estimation, the proposed ground model worked under the assumption that a sufficient number of ground laser points can be observed and approximated by a linear model. Therefore, in highly crowded situations where the ground area is largely occluded by other objects, ground estimation may not be correct or even fail. The detection of objects, especially low-level objects, relies heavily on ground estimation. If ground estimation is wrong or fails, large errors will also occur in object detection. Future work would include the estimation of sensor motion by fusing visual odometry and IMU data, which would further improve the laser profile tracking and object detection. Localization functions using GPS or image-based scene recognition could also be added to improve the usefulness of the system.

Acknowledgments

This research was supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (No. 2012M3C4A7032182).

Authors Contribution

Qing Lin proposed the original idea of the paper and developed the algorithms. Experiments and data analysis were also performed by Qing Lin. Youngjoon Han supervised this work, and provided advices for improving the quality of this work. The manuscript was written by Qing Lin and revised by Youngjoon Han.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. International Agency for Prevention of Blindness. Available online: http://www.iapb.org/ (accessed on 19 July 2014).
2. Dakopoulos, D.; Bourbakis, N.G. Wearable obstacle avoidance electronic travel Aids for blind: A survey. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev*. **2010**, *40*, 25–35.
3. Sainarayanan, G.; Nagarajan, R.; Yaacob, S. Fuzzy image processing scheme for autonomous navigation of human blind. *Appl. Soft Comput*. **2007**, *7*, 257–264.
4. Peng, E.; Peursum, P.; Li, L.; Venkatesh, S. A smartphone-based obstacle sensor for the visually impaired. *Lect. Notes Comput. Sci*. **2010**, *6406*, 590–604.
5. Yu, J.H.; Chung, H.I.; Hahn, H.S. Walking Assistance System for Sight Impaired People Based on a Multimodal Transformation Technique. In Proceedings of the ICROS-SICE International Joint Conference, Fukuoka, Japan, 18–21 August 2009.
6. José, J.; Farrajota, M.; Rodrigues João, M.F.; Hans du Buf, J.M. The Smart Vision local navigation aid for blind and visually impaired persons. *Int. J. Digit. Content Technol. Appl*. **2011**, *5*, 362–375.
7. Lin, Q.; Hahn, H.S.; Han, Y.J. Top-view based guidance for blind people using directional ellipse model. *Int. J. Adv. Robot. Syst*. **2013**, *1*, 1–10.
8. Johnson, L.A.; Higgins, C.M. A Navigation Aid for the Blind Using Tactile-Visual Sensory Substitution. In Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, New York, NY, USA, 31 August–3 September 2006.
9. Dakopoulos, D.; Boddhu, S.K.; Bourbakis, N. A 2D Vibration Array as an Assistive Device for Visually Impaired. In Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, Boston, MA, USA, 14–17 October 2007.
10. Bourbakis, N.; Keefer, R.; Dakopoulos, D.; Esposito, A. A Multimodal Interaction Scheme between a Blind User and the Tyflos Assistive Prototype. In Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence, Dayton, OH, USA, 3–5 November 2008.
11. Bourbakis, N.G.; Keefer, R.; Dakopoulos, D.; Esposito, A. Towards a 2D Tactile Vocabulary for Navigation of Blind and Visually Impaired. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, San Antonio, TX, USA, 11–14 October 2009.
12. Bourbakis, N. Sensing surrounding 3D space for navigation of the blind—A prototype system featuring vibration arrays and data fusion provides a near real-time feedback. *IEEE Eng. Med. Biol. Mag*. **2008**, *27*, 49–55.

13. Meers, S.; Ward, K. A Substitute Vision System for Providing 3D Perception and GPS Navigation via Electro-Tactile Stimulation. In Proceedings of the 1st International Conference on Sensing Technology, Palmerston North, New Zealand, 21–23 November 2005.
14. Gonz ález-Mora, J.L.; Rodr íguez-Hern ández, A.; Rodr íguez-Ramos, L.F.; D íaz-Saco, L.; Sosa, N. Development of a new space perception system for blind people, based on the creation of a virtual acoustic space. *Lect. Notes Comput. Sci*. **1999**, *1607*, 321–330.
15. Dunai, L.; Fajarnes, G.P.; Praderas, V.S.; Garcia, B.D.; Lengua, I.L. Real-time Assistance Prototype—A New Navigation Aid for Blind People. In Proceedings of the 36th Annual Conference on IEEE Industrial Electronics Society, Glendale, AZ, USA, 7–10 November 2010.
16. Rodr´ıguez, A.; Yebes, J.J.; Alcantarilla, P.F.; Bergasa, L.M.; Almaz ´an, J.; Cela, A. Assisting the visually impaired: Obstacle detection and warning system by acoustic feedback. *Sensors* **2012**, *12*, 17477–17496.
17. Pradeep, V.; Medioni, G.; Weiland, J. Robot Vision for the Visually Impaired. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, CA, USA, 13–18 June 2010.
18. Chen, L.; Guo, B.; Sun, W. Obstacle Detection System for Visually Impaired People Based on Stereo Vision. In Proceedings of the 4th International Conference on Genetic and Evolutionary Computing, Shenzhen, China, 13–15 December 2010.
19. Saez, J.M.; Escolano, F. Stereo-Based Aerial Obstacle Detection for the Visually Impaired. In Proceedings of Workshop on Computer Vision Applications for the Visually Impaired, Marseille, France, 18 October 2008.
20. Yuan, D.; Manduchi, R. A Tool for Range Sensing and Environment Discovery for the Blind. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Washinton, DC, USA, 27 June–2 July 2004.
21. Hesch, J.A.; Roumeliotis, S.I. Design and analysis of a portable indoor localization aid for the visually impaired. *Int. J. Robot. Res*. **2010**, *29*, 1400–1415.
22. Baglietto, M.; Sgorbissa, A.; Verda, D.; Zaccaria, R. Human navigation and mapping with a 6DOF IMU and a laser scanner. *Robot. Auton. Syst*. **2011**, *59*, 1060–1069.
23. Fallon, M.F.; Johannsson, H.; Brookshire, J.; Teller, S.; Leonard, J.J. Sensor Fusion for Flexible Human-Portable Building-Scale Mapping. In Proceedings of IEEE International Conference on Intelligent Robots and Systems, Vilamoura, Algarve, Portugal, 7–12 October 2012.
24. Z öllner, M.; Huber, S.; Jetter, H.C.; Reiterer, H. NAVI—A proof-of-concept of a mobile navigational aid for visually impaired based on the Microsoft Kinect. *Lect. Notes Comput. Sci*. **2011**, *6949*, 584–587.
25. Takizawa, H.; Yamaguchi, S.; Aoyagi, M.; Ezaki, N.; Mizuno, S. Kinect, Cane: An Assistive System for the Visually Impaired Based on Three-dimensional Object Recognition. In Proceedings of IEEE International Symposium on System Integration, Fukuoka, Japan, 16–18 December 2012.
26. Khan, A.; Moideen, F.; Lopez, J.; Khoo, W.L.; Zhu, Z. KinDectect: Kinect Detecting Objects. *Lect. Notes Comput. Sci*. **2012**, *7383*, 588–595.

27. Filipea, V.; Fernandesb, F.; Fernandesc, H.; Sousad, A.; Paredese, H.; Barrosof, J. Blind Navigation Support System Based on Microsoft Kinect. In Proceedings of the 4th International Conference on Software Development for Enhancing Accessibility and Fighting Info-exclusion, Douro Region, Portugal, 19–22 July 2012.
28. Brock, M.; Kristensson, P.O. Supporting Blind Navigation Using Depth Sensing and Sonification. In Proceedings of the ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication, Zurich, Switzerland, 8–12 September 2013.
29. Comaniciu, D.; Meer, P. Mean Shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 603–619.
30. Carreira-Perpinan, M.A. Gaussian mean shift is an EM algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 767–776.
31. Zhang, Q.; Pless, R. Extrinsic Calibration of a Camera and Laser Range Finder (Improves Camera Calibration). In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Sendai, Japan, 28 September–2 October 2004.
32. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2003; pp. 138–145.
33. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005.