Perception of Audio Visual Information for Mobile Robot Motion Control Systems

S. Pleshkova[1], Al. Bekiarski[1], Sh. Sehati Dehkharghani[2], Kalina Peeva[2]

[1] Department of Telecommunications, Technical University of Sofia, Bulgaria
{snegpl,aabbv}@tu-sofia.bg
[2] French Language Faculty of Electrical Engineering, Technical University of Sofia, Bulgaria
sh.sehati@gmail.com kala_peeva@yahoo.com

Abstract. Motion is the main characteristic of intelligent mobile robots. There exist a lot of methods and algorithms for mobile robots motion control. These methods are based on different principles, but the results from these methods must leads to one final goal - to provide a precise mobile robot motion control with clear orientation in the area of robot perception and observation. First in the proposed chapter are outlined the mobile robot audio and visual systems with the corresponding audio (microphone array) and video (mono, stereo or thermo cameras) sensors, accompanied with laser range finder sensor. The audio and video information captured from the sensors is used in the perception audio visual model proposed to perform joint processing of audio visual information and to determine the current mobile robot position (current space coordinates) in the area of robot perception and observation. The captured from audio visual sensors information is estimated with the suitable algorithms developed for speech and image quality estimation to apply the preprocessing methods for increasing the quality and to minimizing the errors of mobile robot position calculations. The current space coordinates determined from laser range finder are used as a supplementary information of mobile robot position, for error calculation and for comparison with the results from audio visual mobile robot motion control. In the development of the mobile robot perception audio visual model are used: method (RANSAC - Random Sample Consensus) for estimate parameters of a mathematical model from a set of observed audio visual coordinate data; method (DOA - Direction of Arrival) for sound source direction localization with microphone array of speaker sending voice commands to the mobile robot; method for speech recognition of the voice command sending from the speaker to the robot.

The current mobile robot position calculated from joint usage of perceived audio visual information is used in appropriate algorithms for mobile robot navigation, motion control and objects tracking: map based or map less methods, path planning and obstacle avoidance, simultaneous localization and mapping (SLAM), data fusion, etc.

The error, accuracy and precision of the proposed mobile robot motion control with perception of audio visual information are analyzed and estimated from the results of the numerous experimental tests presented at the end of this chapter. The experiments are carried out mainly with simulations of the algorithms listed above, but are trying also parallel computing methods in implementation of the developed algorithms to reach real time robot navigation an motion control using perceived audio visual information from the mobile robot audio visual sensors.

Keywords. Audio visual perception, Mobile robot motion control, Audio visual object tracking

Contents:

1 Introduction. Audio visual robot perception

Robot perception is an important characteristic of all modern intelligent robots [1] closely related to the human perception [2]. Although the robot perception tries to copy human perception system, there are significant differences between human and robot perception. These differences are not only in the hardware and software robot perception system realization. Generally these differences are in understanding and in

precision of modeling the human perception using mathematical interpretation or heuristic interpretation of visual information from the environments around the mobile robots. There are a lot of scientific publications and articles trying to solve the general robot perception problem mostly related and compared them with the same characteristics of the human perceptions [3], [4], [5], [6]. The main advantages from these articles are the conclusions that it is necessary to have a general representative robot perception model containing all existing human perception characteristics and applying this general model for solving concrete more often practical robot application tasks. There exist a large number of examples of robot applications, where the robot perception models help to develop the effective algorithms in wide range of robot applications from robot manipulators [7], [8], [9] to mobile robots [10], [11], [12], [13] and humanoid like robots [14], [15]. In this chapter the attention is focused to mobile robot perception and especially for mobile robot motion control. The task of mobile robot motion control is well presented in scientific literature [16], [17], [18] and also there are the applications strictly directed only to audio robot perception [19] and only to visual robot perception [20]. The proposed audio visual mobile robot perception system and algorithms for mobile robot motion control in this chapter are based on the condition to perform joint processing of audio visual information and to determine the current mobile robot position (current space coordinates) in the area of robot perception and observation.

In this work[1], an audio-visual system is proposed for mobile robot motion control based on audio-visual and range information perceived from robot sensors (microphone array, video camera, and laser rangefinder). Robot motion is controlled through speech commands and EKF-SLAM is applied for robot navigation. In EKF-SLAM, the environment landmarks are vertical edges, perceived from the camera image, associated to corners, perceived from the 2D laser rangefinder. This way of modeling of the environment has the advantage that because there is not high feature clutter, the problem of quadratic complexity of the EKF-SLAM is solved to a great extent. On the other hand it constraints the proposed system to be applicable only in structured indoor environments, containing enough vertical edges. The general system is presented in Section 2. Sensor calibration and robot navigation based on EKF-SLAM are explained in detail in Section 3 and Section 4, respectively. In Section 5 an algorithm is presented for quality estimation of perceived speech information. Experimental results in Section 6 show the functionality of the proposed system. The conclusion in Section 7 puts an end to this paper.

[1] The proposed system for mobile robot motion control through speech commands is based on parts of the researches done towards the PhD thesis "Development of Methods and Algorithms for Audio-Visual Mobile Robot Motion Control ", conducted at The French Language Faculty of Electrical Engineering, Technical University of Sofia, Bulgaria [21].

2 Mobile robot audio and visual perception system with corresponding audio visual sensors and additional laser range finder sensor

The general system for mobile robot navigation based on its perceptions from its environment is presented in Fig. 1. It is assumed that the mobile robot perceives audio-visual information from its environment through a microphone and a video camera and range information through a laser rangefinder. The path of the robot is planned based on the perceived audio information. In other words it is possible to command the robot to navigate within its environment through speech commands. It is described in more detail in Sections 4.2 and 4.3. Robot navigation is based on EKF-SLAM, in which the environment is modeled based on perceived visual and range information from the camera and laser rangefinder. Therefore, the sensors are calibrated in order to compensate the systematic errors and also to calculate the relative position of sensors to be able to associate visual perceptions with range information.

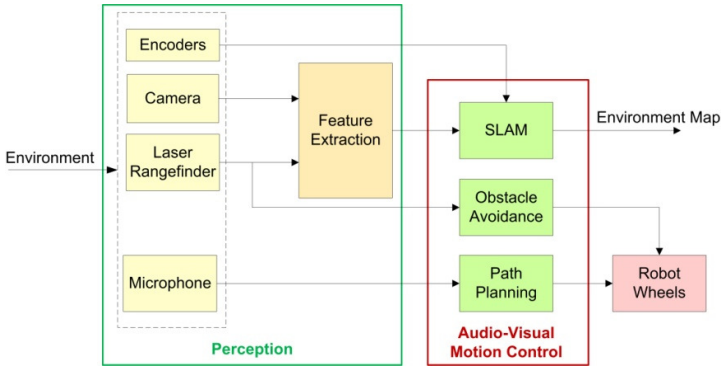


Fig. 1. General system for mobile robot navigation within its environment based on perceived audio-visual and range information

3 Sensor calibration using mobile robot visual and range perceptions

The first preparatory step for any system, which consists of several sensors is the calibration step [22]. Sensor calibration is performed in order to compensate the systematic errors of sensor measurements and being able to transform object coordinates from the world reference frame to the local frame of the sensor and vice versa. By geometric calibration of the camera, its intrinsic parameters are obtained which are used for compensation of lens distortions. Intrinsic camera calibration method is discussed in Section 3.1.

After intrinsic calibration of the camera, laser rangefinder is calibrated with it extrinsically. In this way the relative position of the sensors are obtained. Subsequently, it is easy to find correspondence between the data provided by each of them.

Therefore, in order to achieve precise results from the proposed system, camera parameters are first computed by geometric calibration of the camera. Then, the 2D laser rangefinder is calibrated extrinsically with the camera. In this way, compensation of systematic errors is ensured and measurements can be modeled by a Gaussian distribution containing uncertainty (three standard deviation) caused by random errors. It is also possible to find correspondence between vertical edges in the image received from the camera and corners in laser data. Thus, each feature is presented by its bearing extracted from visual data and its corresponding range extracted from laser data.

3.1 Geometric video camera calibration from perceived visual information of mobile robot

Geometric camera calibration method provides camera parameters used for the transformation of the object coordinates from the 3D world reference frame to 2D image frame and vice versa based on a set of images captured by the camera from a test object with a unique pattern from different positions (with varying angle and depth). The most popular pattern is a printed chessboard pattern. It is important that the pattern produces distinct and well defined corners in the set of images used for camera calibration.

The intrinsic and extrinsic parameters of the camera are determined based on an ideal pinhole camera model. Intrinsic camera parameters include camera's focal length, the principal point, lens distortions (tangential and radial) and scaling factors (for transformation from 3D metric world reference frame to 2D metric image frame and from metric units to pixels). And, extrinsic parameters are rotation matrix and translation vector of the object reference frame with respect to the camera reference frame. Pixels are assumed to be rectangular (zero skew).

The coordinate systems used for camera calibration procedure is presented in Fig. 2 [22].

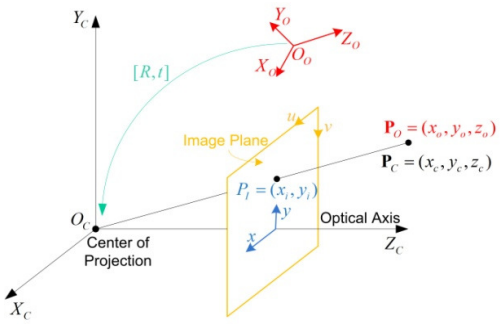


Fig. 2. Coordinate systems used in geometric camera calibration procedure

$\mathbf{P}_O = (x_o, y_o, z_o)$ and $\mathbf{P}_C = (x_c, y_c, z_c)$ are object coordinates with respect to its local frame and camera frame, respectively. $\mathbf{P}_I = (u_i, v_i)$ represents object coordinates in the image plane in pixels.

Assuming that the object local frame with respect to the camera reference frame is represented by a 3×3 rotation matrix, $\mathbf{R}$ and a 3×1 translation vector, $\mathbf{t}$, object coordinates in the camera frame are:

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \underbrace{\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}}_{R} \begin{bmatrix} x_o \\ y_o \\ z_o \end{bmatrix} + \underbrace{\begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix}}_{t} \tag{1}$$

The coordinates of the corresponding point projected to the image plane are calculated using Eq. (2).

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} = \frac{f}{z_c} \begin{bmatrix} x_c \\ y_c \end{bmatrix}; \tag{2}$$

where

f is the focal length of the camera.

Then, the projected point in the image plane is represented in pixels (u_i', v_i'):

$$\begin{bmatrix} u_i' \\ v_i' \end{bmatrix} = \begin{bmatrix} D_u s_u x_i \\ D_v y_i \end{bmatrix} + \begin{bmatrix} u_0 \\ v_0 \end{bmatrix}; \tag{3}$$

where

s_u - the scale factor;

D_u, D_v - coefficients for conversion from metric units to pixels;

$\begin{bmatrix} u_0 & v_0 \end{bmatrix}^T$ - principal point.

The pinhole camera assumption is an ideal assumption. Radial and tangential lens distortions are added to the ideal model in order to correct this assumption. Here, only two coefficients are considered for each distortion. The radial and tangential distortions are modeled by Eqs. (4) and (5), respectively.

$$\begin{bmatrix} \Delta u_i^{(r)} \\ \Delta v_i^{(r)} \end{bmatrix} = \begin{bmatrix} x_i (k_1 r_i^2 + k_2 r_i^4) \\ y_i (k_1 r_i^2 + k_2 r_i^4) \end{bmatrix}; \tag{4}$$

where

k_1, k_2 are radial distortion coefficients, and $r_i = \sqrt{x_i^2 + y_i^2}$.

$$\begin{bmatrix} \Delta u_i^{(t)} \\ \Delta v_i^{(t)} \end{bmatrix} = \begin{bmatrix} 2p_1 x_i y_i + p_2 (r_i^2 + 2x_i^2) \\ p_1 (r_i^2 + 2y_i^2) + 2p_2 x_i y_i \end{bmatrix}; \tag{5}$$

where

p_1, p_2 are tangential distortion coefficients.

Therefore, the general camera calibration model is obtained by correcting the pinhole model by combining the pinhole model and radial and tangential distortions (Eq. 6).

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} = \begin{bmatrix} \alpha_u \tilde{u}_i \\ \alpha_v \tilde{v}_i \end{bmatrix} + \begin{bmatrix} u_0 \\ v_0 \end{bmatrix} = \begin{bmatrix} D_u s_u (x_i + \Delta u_i^{(r)} + \Delta u_i^{(t)}) \\ D_v (y_i + \Delta v_i^{(r)} + \Delta v_i^{(t)}) \end{bmatrix} + \begin{bmatrix} u_0 \\ v_0 \end{bmatrix} \tag{6}$$

where $(\tilde{u}_i, \tilde{v}_i)$ are distorted coordinates.

The camera calibration parameters can be estimated linearly using Direct Linear Transform (DLT) method [23]. In this approach nonlinear radial and tangential distortions are ignored and transformation from object local frame to image frame is assumed to be linear using the homogeneous 3×4 matrix, M.

$$\begin{bmatrix} u_i w_i \\ v_i w_i \\ w_i \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \end{bmatrix} \begin{bmatrix} x_o \\ y_o \\ z_o \\ 1 \end{bmatrix} \tag{7}$$

By eliminating the depth value, w_i, for each control point (x_j, y_j, z_j) ; $j = 1, 2, ..., N$, Eq. (8) is valid.

$$\mathbf{L}_j \mathbf{m} = 0 \; ; \; j = 1, 2, ..., N \tag{8}$$

where

$$\mathbf{m} = \begin{bmatrix} m_{11}, m_{12}, m_{13}, m_{14}, m_{21}, m_{22}, m_{23}, m_{24}, m_{31}, m_{32}, m_{33}, m_{34} \end{bmatrix}^T$$

and

$$\mathbf{L}_j = \begin{bmatrix} x_j & y_j & z_j & 1 & 0 & 0 & 0 & 0 & -x_j u_j & -y_j u_j & -z_j u_j & -u_j \\ 0 & 0 & 0 & 0 & x_j & y_j & z_j & 1 & -x_j v_j & -y_j v_j & -z_j v_j & -v_j \end{bmatrix}$$

By replacing (u_j, v_j) with the coordinates of the observed points (U_j, V_j), the values of $m_{11},..m_{34}$ can be estimated using the least squares method. In order to avoid singularities, in [24] is proposed to use the constraint $m_{31}^2 + m_{32}^2 + m_{33}^2 = 1$.

Figure 3 shows the main steps of the geometric camera calibration algorithm [22]. In the proposed method a sequence of images of the chessboard pattern is captured by the camera from different positions with varying depth and angle. Then, the user selects the extreme grid corners for each image and inputs some geometrical information about the dimensions of the grid cells. These values are used for corner extraction initialization. The coordinates of the points P_i in the image plane are the location of all corners detected in each observed image.

In this stage of the calibration procedure, it is assumed that during image observation only Gaussian noise is present and systematic measurement noise is compensated. Therefore, camera calibration parameters are computed by minimizing the least square error between the observed coordinates and the coordinates computed based on the calibration model presented in Eq. (6). Considering N corner observations $\{(U_1, V_1), \ldots, (U_N, V_N)\}$, least squares method is used to minimize Eq. (9).

$$E^2 = \sum_{i=1}^{N} (U_i - u_i)^2 + \sum_{i=1}^{N} (V_i - v_i)^2 \tag{9}$$

Because the calibration model is nonlinear, calibration parameters are estimated iteratively by minimizing Eq. (9) using the Levenberg–Marquardt algorithm (LMA) [25]. In order to avoid the local minimum problem during the iterative optimization process, the initial values of the parameters are computed using the DLT method.
Computed camera calibration parameters are used for image correction. Table 1 presents the image correction algorithm. First, a 40×40 grid with tie-points (x_i, y_i) is generated, covering the entire image. Distorted coordinates, $(\tilde{u}_i, \tilde{v}_i)$ of the corresponding tie-points are calculated.

Then, parameters $a_1, \ldots, a_8$ of Eq. (10) are estimated iteratively using the least squares method in order to calculate the undistorted coordinates.

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} = \frac{1}{N} \begin{bmatrix} \tilde{u}_i(1 + a_1 r_i^2 + a_2 r_i^4) + 2a_3 \tilde{u}_i \tilde{v}_i + a_4 (r_i^2 + 2\tilde{u}_i^2) \\ \tilde{v}_i(1 + a_1 r_i^2 + a_2 r_i^4) + a_3 (r_i^2 + 2\tilde{v}_i^2) + 2a_4 \tilde{u}_i \tilde{v}_i \end{bmatrix}; \tag{10}$$

where

$N = (a_5 r_i^2 + a_6 \tilde{u}_i + a_7 \tilde{v}_i + a_8) r_i^2 + 1$; and $r_i^2 = \tilde{u}_i^2 + \tilde{v}_i^2$.

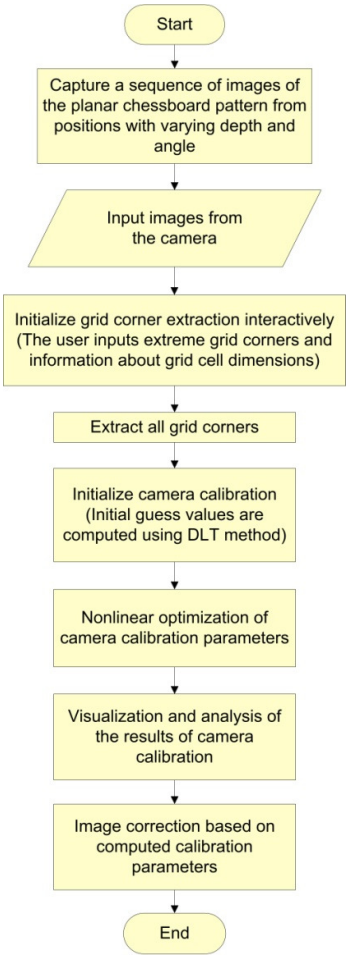


Fig. 3. Geometric camera calibration algorithm

Once the parameters are estimated, Eq. (10) can be employed for the computation of the corresponding undistorted coordinates. Actual coordinates of the points of the image are calculated by interpolating the computed distorted and corresponding undistorted results.

Table 1. Image correction algorithm

Image correction algorithm
1. Generate a 40×40 grid with distorted and undistorted tie-points (x_i, y_i) and $(\tilde{u}_i, \tilde{v}_i)$, covering the entire image. 2. Calculate the corresponding distorted coordinates, $(\tilde{u}_i, \tilde{v}_i)$. 3. Estimate parameters $a_1, \ldots, a_8$ for calculation of the undistorted coordinates iteratively using the least squares method. 4. Compute the corrected undistorted coordinates, (x_i, y_i) based on the estimated parameters. 5. Calculate all actual coordinates of the image by interpolation based on $(\tilde{u}_i, \tilde{v}_i)$ and the new (x_i, y_i).

3.2 Camera-laser rangefinder extrinsic calibration from information of 2D laser rangefinder, visual sensor and geometric video camera Calibration

After intrinsic camera calibration, extrinsic calibration of the 2D laser rangefinder and the camera is performed in order to calculate the relative position of the camera local frame with regard to the laser local frame by providing the translation vector, $\mathbf{t}_{cl}$ and the rotation matrix, $\mathbf{R}_{cl}$. As a result, point $\mathbf{P}_c$ in the camera frame can be corresponded with point $\mathbf{P}_l$ in the laser frame using Eq. (11).

$$\mathbf{P}_l = \mathbf{R}_{cl}\mathbf{P}_c + \mathbf{t}_{cl} \tag{11}$$

The calibration is based on observing the same test pattern by both of the sensors from different positions. The position of the test pattern in each observation is obtained based on parameters achieved in camera calibration in previous step, and the line corresponding to the board in laser data is extracted iteratively by minimizing the re-projection error. In order to be able to extract the planar chessboard pattern from laser data in each observation, the planar chessboard has to be moved in an almost static environment for each observation. Therefore, one of the constraints of extracting the board line in each image is that it is a straight line which changes position in each observation. The other constraint comes from considering the fact that the points belonging to the board line in laser data must lie on the calibration plane, extracted from camera calibration. In other words, assuming that the calibration plane, N is on the plane $z = 0$, and is presented by translation vector, t and rotation matrix, R, provided by camera calibration (Eq. 12), coordinates of each point $\mathbf{P}_l$ of the board line in laser data must be on plane N (Eq. 13).

$$\mathbf{N} = -\mathbf{R}_3(\mathbf{R}_3^T \cdot \mathbf{t}_0); \tag{12}$$

where $\mathbf{R}_3$ is the third column of the rotation matrix R, and $\mathbf{t}_0$ is the center of the camera in the world frame.

$$\mathbf{N} \cdot \mathbf{R}_{cl}^{-1}(\mathbf{P}_l - \mathbf{t}_{cl}) = |\mathbf{N}|^2 \tag{13}$$

It is evident from Eq. (13) that N is normal to the calibration board and its magnitude is equal to the distance from the center of the camera to the calibration board.

Assuming that all board lines in laser data are on the plane $y = 0$, $\mathbf{R}_{cl}$ and $\mathbf{t}_{cl}$ can be estimated by minimizing iteratively the error, which is defined by the sum of Euclidian distance of laser points from the calibration plane, using Levenberg-Marquardt method. Outliers can also be removed considering the first constraint. Therefore, assuming the two described constraints, the translation vector and rotation matrix are computed iteratively by minimizing the error in the re-projection of the board line on the camera image.

4 Navigation of mobile robot from perception of audio visual information

As is already mentioned in Section 1, the robot is going to follow speech commands in structured unknown indoor environments. Therefore, robot navigation within its environment is based on EKF-SLAM. It is described briefly in Section 4.1. Robot path planning, presented in Section 4.2, depends on the recognized speech command.

4.1 Robot navigation based on EKF-SLAM

A robot placed in an unknown environment can concurrently build the map of its surrounding environment while localizing itself within it performing Simultaneous Localization and Mapping (SLAM). In general the probabilistic definition of SLAM is: At each time t, given the control input (obtained from encoders) $\mathbf{u}_t$, and a set of landmark observations (sensor measurements) z_t, the joint posterior density of the robot state and the landmark locations has the following distribution:

$$P(\mathbf{x}_t, \mathbf{M} \mid \mathbf{z}_t, \mathbf{u}_t) \tag{14}$$

Using Bayes rule the posterior can be written as:

$$P(\mathbf{x}_t, \mathbf{M} \mid \mathbf{z}_t, \mathbf{u}_t) = \eta P(\mathbf{z}_t \mid \mathbf{x}_t, \mathbf{M}) P(\mathbf{x}_t, \mathbf{M} \mid \mathbf{z}_{t-1}, \mathbf{u}_t) \tag{15}$$

By applying the Theorem of Total Probability [26] and then the definition of the conditional probability to Eq. (15), the posterior is described as:

$$P(\mathbf{x}_t, \mathbf{M} \mid \mathbf{z}_t, \mathbf{u}_t) = \eta P(\mathbf{z}_t \mid \mathbf{x}_t, \mathbf{M}) \int P(\mathbf{x}_t, \mathbf{M} \mid \mathbf{x}_{t-1}, \mathbf{z}_{t-1}, \mathbf{u}_t) P(\mathbf{x}_{t-1} \mid \mathbf{z}_{t-1}, \mathbf{u}_t) d\mathbf{x}_{t-1}$$

$$= \eta P(\mathbf{z}_t \mid \mathbf{x}_t, \mathbf{M}) \int P(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{u}_t) P(\mathbf{M} \mid \mathbf{x}_{t-1}, \mathbf{z}_{t-1}, \mathbf{u}_t) P(\mathbf{x}_{t-1} \mid \mathbf{z}_{t-1}, \mathbf{u}_t) d\mathbf{x}_{t-1}$$

$$= \eta P(\mathbf{z}_t \mid \mathbf{x}_t, \mathbf{M}) \int P(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{u}_t) P(\mathbf{x}_{t-1}, \mathbf{M} \mid \mathbf{z}_{t-1}, \mathbf{u}_t) d\mathbf{x}_{t-1} \quad (16)$$

The resultant recursive equation shows that the SLAM posterior is a function of the measurement model $P(\mathbf{z}_t \mid \mathbf{x}_t, \mathbf{M})$, the motion model $P(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{u}_t)$, and the SLAM posterior at time $t-1$.

The Extended Kalman Filter (EKF) is the most common estimation of the SLAM posterior, which represents it as a high-dimensional, multivariate Gaussian parameterized by a mean $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$ [27,28].

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{X}} \\ \boldsymbol{\mu}_{\mathbf{M}} \end{bmatrix} = \begin{bmatrix} \mu_{x_r} \\ \mu_{y_r} \\ \mu_{\theta r} \\ \boldsymbol{\mu}_{L1} \\ \vdots \\ \boldsymbol{\mu}_{L_N} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} & \boldsymbol{\Sigma}_{\mathbf{XM}} \\ \boldsymbol{\Sigma}_{\mathbf{MX}} & \boldsymbol{\Sigma}_{\mathbf{M}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} & \boldsymbol{\Sigma}_{\mathbf{XL_1}} & \cdots & \boldsymbol{\Sigma}_{\mathbf{XL_N}} \\ \boldsymbol{\Sigma}_{\mathbf{L_1X}} & \boldsymbol{\Sigma}_{\mathbf{L_1L_1}} & \cdots & \boldsymbol{\Sigma}_{\mathbf{L_1L_N}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{\mathbf{L_NX}} & \boldsymbol{\Sigma}_{\mathbf{L_NL_1}} & \cdots & \boldsymbol{\Sigma}_{\mathbf{L_NL_N}} \end{bmatrix}$$

(17)

The system state consists of the set of landmark locations M (where the coordinates are with respect to the world reference frame) and the robot state. Figure 4 demonstrates the construction of the system state in brief [21].

The non-linear motion model $\mathbf{x}_t = f(\mathbf{x}_{t-1}, \mathbf{u}_t) + \mathbf{w}_t$ ($\mathbf{w}_t$ is the zero-mean Gaussian system noise with the covariance $\mathbf{Q}_t$), and measurement model $\mathbf{z}_t = h(\mathbf{x}_{t-1}) + \mathbf{r}_t$ ($\mathbf{r}_t$ is the Gaussian measurement error with the covariance $\mathbf{R}_t = \begin{bmatrix} \sigma_r^2 & 0 \\ 0 & \sigma_\phi^2 \end{bmatrix}$), are linearized around the most likely system state $\boldsymbol{\mu}_{t-1}$, using Taylor Expansion: $g \approx g + g'$ where, g' is the Jacobian of the function with respect to its variables.

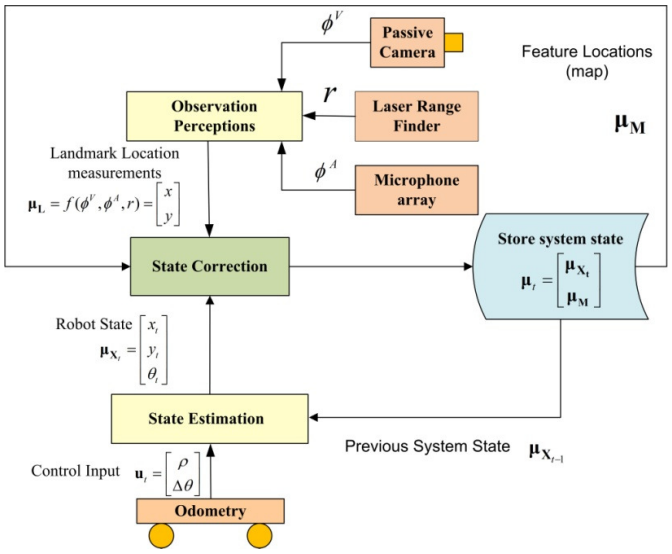


Fig. 4. Developed schema for system state construction in EKF-SLAM

Considering the above assumptions, the EKF-SLAM is a recursive algorithm that can be divided into two main steps: state estimation (prediction) and state update (correction).

1. State estimation (Prediction)

The estimated state vector and covariance matrix are calculated from the previous state and covariance, and the control input:

$$\bar{\mathbf{x}}_t = f(\mathbf{x}_{t-1}, \mathbf{u}_t), \qquad \bar{\boldsymbol{\Sigma}}_t = \mathbf{F}_{\mathbf{x},t}\boldsymbol{\Sigma}_{t-1}\mathbf{F}_{\mathbf{x},t}^T + \mathbf{F}_{\mathbf{u},t}\mathbf{Q}_t\mathbf{F}_{\mathbf{u},t}^T, \tag{18}$$

where, $\mathbf{F}_{\mathbf{x},t} = \left.\dfrac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}=\boldsymbol{\mu}_{t-1}}$ is the Jacobian of the state transition function f with respect to the robot state.

Figure 5 demonstrates the motion model of the two-wheeled nonholonomic mobile robot. The predicted system state and covariance, considering the odometry model $\mathbf{u}_t = \boldsymbol{\mu}_{\mathbf{u}_t} + \mathbf{w}_t$, are:

$$\bar{\boldsymbol{\mu}} = \begin{bmatrix} \bar{\boldsymbol{\mu}}_{\mathbf{X}} \\ \boldsymbol{\mu}_{\mathbf{M}} \end{bmatrix}, \ \bar{\boldsymbol{\Sigma}} = \begin{bmatrix} \bar{\boldsymbol{\Sigma}}_{\mathbf{X}} & \bar{\boldsymbol{\Sigma}}_{\mathbf{XM}} \\ \bar{\boldsymbol{\Sigma}}_{\mathbf{MX}} & \boldsymbol{\Sigma}_{\mathbf{M}} \end{bmatrix} \tag{19}$$

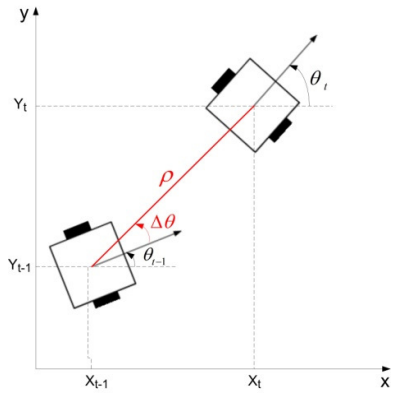


Fig. 5. Motion model of a two-wheeled nonholonomic mobile robot

The only time variant part of the system state is the robot state. Therefore, $\Sigma_M = \overline{\Sigma}_M$, and $\boldsymbol{\mu}_M = \overline{\boldsymbol{\mu}}_M$ (the estimated and updated map features with regard to the world reference frame are the same).

The robot state and system covariance are predicted as follows:

$$\overline{\boldsymbol{\mu}}_{\mathbf{X}_t} = \boldsymbol{\mu}_{\mathbf{X}_{t-1}} + \boldsymbol{\mu}_{\mathbf{u}_t} = \begin{bmatrix} \overline{x}_t \\ \overline{y}_t \\ \overline{\theta}_t \end{bmatrix} = \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ \theta_{t-1} \end{bmatrix} + \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta\theta \end{bmatrix} = \begin{bmatrix} x_{t-1} + \rho\cos(\theta_{t-1} + \Delta\theta) \\ y_{t-1} + \rho\sin(\theta_{t-1} + \Delta\theta) \\ \theta_{t-1} + \Delta\theta \end{bmatrix} \quad (20)$$

$$\overline{\boldsymbol{\Sigma}}_{\mathbf{X}_t} = \mathbf{F}_{\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{X}_{t-1}}\mathbf{F}_{\mathbf{x}}^T + \mathbf{F}_{\mathbf{u}}\mathbf{Q}_t\mathbf{F}_{\mathbf{u}}^T \;,\; \overline{\boldsymbol{\Sigma}}_{\mathbf{X}_t\mathbf{M}} = \mathbf{F}_{\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{X}_{t-1}\mathbf{M}} \;,\; \overline{\boldsymbol{\Sigma}}_{\mathbf{M}\mathbf{X}_t} = \overline{\boldsymbol{\Sigma}}_{\mathbf{X}_t\mathbf{M}}^T \quad (21)$$

where $\boldsymbol{F_x}$ and $\boldsymbol{F_u}$ are Jacobians of the state transition function with respect to the robot state and control input, respectively:

$$\mathbf{F}_{\mathbf{x}} = \begin{bmatrix} 1 & 0 & -\rho\sin(\theta_{t-1} + \Delta\theta) \\ 0 & 1 & \rho\cos(\theta_{t-1} + \Delta\theta) \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\Delta y \\ 0 & 1 & \Delta x \\ 0 & 0 & 1 \end{bmatrix} \quad (22)$$

$$\mathbf{F}_{\mathbf{u}} = \begin{bmatrix} \cos(\theta_{t-1} + \Delta\theta) & -\rho\sin(\theta_{t-1} + \Delta\theta) \\ \sin(\theta_{t-1} + \Delta\theta) & \rho\cos(\theta_{t-1} + \Delta\theta) \end{bmatrix} \quad (23)$$

2. State update (Correction)

For each keypoint observation $\mathbf{z}_t = \begin{bmatrix} r_t \\ \phi_t \end{bmatrix}$, data association is performed. The extracted keypoint can be associated to a landmark in the database or will be considered to be added to the database if it is not associated to any known landmark. In the latter case, if the landmark is observed at least a specific number of times, it will be added to the database. Assuming that the location of the observed landmark in the map is $\bar{\boldsymbol{\mu}}_{L_i} = \begin{bmatrix} \bar{x}_i \\ \bar{y}_i \end{bmatrix}$, expected measurement is obtained as follows:

$$\boldsymbol{\delta} = \begin{bmatrix} \delta_x \\ \delta_y \end{bmatrix} = \begin{bmatrix} \bar{x}_t - \bar{x}_i \\ \bar{y}_t - \bar{y}_i \end{bmatrix} \tag{24}$$

$$\lambda = \delta^T \delta \tag{25}$$

$$h(\bar{\boldsymbol{\mu}}_t) = \begin{bmatrix} \sqrt{\lambda} \\ \arctan(\frac{\delta_y}{\delta_x}) - \bar{\theta}_t \end{bmatrix} \tag{26}$$

The Jacobian of the measurement model with respect to robot state, $\mathbf{H}_{\mathbf{x},t}$ is:

$$\mathbf{H}_{\mathbf{x},t} = \frac{1}{\lambda} \begin{bmatrix} -\sqrt{\lambda}\delta_x & -\sqrt{\lambda}\delta_y & 0 & 0 & \dots & 0 & \sqrt{\lambda}\delta_x & \sqrt{\lambda}\delta_y & 0 & \dots & 0 \\ \delta_y & -\delta_x & -\lambda & 0 & \dots & 0 & -\delta_y & \delta_x & 0 & \dots & 0 \end{bmatrix}$$

$$\underbrace{\qquad\qquad}_{2i-2} \qquad\qquad \underbrace{\qquad\qquad}_{2N-2i} \tag{27}$$

The Kalman gain, and the system state and covariance are updated as follows:

$$\mathbf{K}_t = \bar{\boldsymbol{\Sigma}}_t \mathbf{H}_{\mathbf{x},t}^T \underbrace{(\mathbf{H}_{\mathbf{x},t} \bar{\boldsymbol{\Sigma}}_t \mathbf{H}_{\mathbf{x},t}^T + \mathbf{R}_t)}_{\text{Innovation Covariance}}^{-1} \tag{28}$$

$$\boldsymbol{\mu}_t = \bar{\boldsymbol{\mu}}_t + \mathbf{K}_t (\mathbf{z}_t - h(\bar{\boldsymbol{\mu}}_t)) \tag{29}$$

$$\boldsymbol{\Sigma}_t = (\mathbf{I} - \mathbf{K}_t \mathbf{H}_{\mathbf{x},t}) \bar{\boldsymbol{\Sigma}}_t \tag{30}$$

Landmark update (Augmentation). If a keypoint is stable enough to be added to the map, the state vector and the covariance matrix are updated to contain the new landmark.

$$\boldsymbol{\mu}_t = \begin{bmatrix} \boldsymbol{\mu}_t \\ x_{N+1} \\ y_{N+1} \end{bmatrix}, \boldsymbol{\Sigma}_t = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} & \boldsymbol{\Sigma}_{\mathbf{XM}} & \boldsymbol{\Sigma}_{\mathbf{X}}^T \mathbf{J}_{\mathbf{X}}^T \\ \boldsymbol{\Sigma}_{\mathbf{XM}}^T & \boldsymbol{\Sigma}_{\mathbf{M}} & \boldsymbol{\Sigma}_{\mathbf{XM}}^T \mathbf{J}_{\mathbf{X}}^T \\ \mathbf{J}_{\mathbf{X}} \boldsymbol{\Sigma}_{\mathbf{X}} & \mathbf{J}_{\mathbf{X}} \boldsymbol{\Sigma}_{\mathbf{XM}} & \mathbf{J}_{\mathbf{X}} \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{J}_{\mathbf{X}}^T + \mathbf{J}_{\mathbf{z}} \mathbf{R}_t \mathbf{J}_{\mathbf{z}}^T \end{bmatrix} \tag{31}$$

where $\mathbf{J}_{\mathbf{X}}$ and $\mathbf{J}_{\mathbf{z}}$ represent Jacobians of landmark prediction with respect to robot state and measurement variables, respectively.

$$\mathbf{J}_{\mathbf{X}} = \begin{bmatrix} 1 & 0 & -\rho \sin(\theta_{t-1} + \Delta\theta) \\ 0 & 1 & \rho \cos(\theta_{t-1} + \Delta\theta) \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\Delta y \\ 0 & 1 & \Delta x \end{bmatrix} \tag{32}$$

$$\mathbf{J}_{\mathbf{z}} = \begin{bmatrix} \cos(\theta_{t-1} + \Delta\theta) & -\rho \sin(\theta_{t-1} + \Delta\theta) \\ \sin(\theta_{t-1} + \Delta\theta) & \rho \cos(\theta_{t-1} + \Delta\theta) \end{bmatrix} \tag{33}$$

Landmark extraction. Movement of the robot is considered to be two-dimensional. A good map of the environment can be obtained using both visual and laser data. Corners in laser data associated with vertical edges in the camera image are landmarks of the environment. The laser rangefinder provides accurate distance data to the edges, while accurate bearing information is computed from the images acquired by the camera. In order to extract vertical edges in the images provided by visual sensor, Hough transform [29] is employed to the resultant binary image of Canny edge detector [30]. Corners are extracted from raw laser information, and only the ones that can be corresponded to a vertical edge in the image are considered as map features. Since the laser rangefinder and camera are calibrated extrinsically, vertical edges in image can be easily corresponded to corner points in laser data.

Data Association. Data association finds correspondence between the current landmark database and the new observations. In this paper, observation is based on gated nearest-neighbor approach, in which each matching between sensor observations and map features is considered independently. In this approach, correlation between measurement prediction errors is ignored. This will cause problems by accepting bad data associations in high clutter or when robot error increases. Because map features are based on perceptions from two sensors and only corners in the laser data which correspond to vertical edges in the camera image are considered as map features, it is assumed that there is not high feature clutter. This also deals with the problem of quadratic complexity of the EKF-SLAM.

The feature database is a table, in which each landmark is defined as $\mathbf{L}_i = (x_i, y_i, h_i, m_i)$, where x_i and y_i are landmark locations, h_i and m_i show the number of hits and misses of each keypoint during data association. The first time a

keypoint is extracted, $h_i = 1$ and $m_i = 0$. If the minimum probability is above some fixed threshold, the observation is considered for addition as a new landmark (h_i will be incremented). On the other hand, if a landmark is predicted to be in the field of view of the camera but is not associated with any observation, it is missed (m_i is incremented). A keypoint that is hit more than a specific number of times will be added to the map, and a landmark in the map that is missed a specific number of times is suppressed from the map.

4.2 Path planning based on perceived audio information

The robot is controlled by speech commands [21]. Each speech command is a set of isolated pre-trained words, like: "STOP MOVING, COME HERE, TURN LEFT, TURN RIGHT, STOP AT, FOLLOW ME, ..." and some numbers. All of these commands are saved in a database in the memory of the computer and for each of the commands, there is a function for interpreting the corresponding command for the robot. Table 2 shows the list of the commands and the corresponding robot actions.

Table 2. List of commands and the corresponding robot actions

The command	*The action*
TURN LEFT	The robot turns left 90 degrees.
TURN RIGHT	The robot turns right 90 degrees.
STOP MOVING	The robot stops moving.
CONTINUE x* METERS	The robot continues *x* meters straight with its current direction.
STOP AT x* AND y*	The robot moves towards the goal point (*x*,*y*) and stops there.
FOLLOW WALL	The robot keeps moving parallel to the wall on its right side while keeping a distance of 0.5 meters from it.
COME HERE	The robot moves towards the calculated location of sound source and stops at a distance of 0.5 meters from it.
FOLLOW ME	The robot rotates towards the calculated sound source direction and tracks the displacements of the human detected in that direction in the camera image.

* *x* and *y* are in meters.

In the proposed system for audio-visual mobile robot motion control, the program for Windows Speech Recognition application is used, which is based on HMM. Once a speech signal is received by the microphones, the speech-to-text program, which is being run asynchronously in parallel to the proposed system, writes the detected

speech signal in a specific predefined text file, and the sound source direction in another text file. This program is independent from the main system, is executed in parallel, and uses Speech Application Programming Interface (SAPI) to write the received speech signals in a text file, that is going to be read in the program of the main system. It also writes the calculated sound source location (described in Section 4.3), in another text file.

The text file, that contains information about the sound source location, is only used (read) in the main program if it is detected that the received command is "COME" or "FOLLOW ME". Otherwise, the robot follows the received command even if no human is present in the environment.

As is demonstrated in Fig. 6 [21], the robot waits for speech input by reading the text file in which the received command is written. If the text file is not empty, the speech command-based path planner reads the text and compares it with the command words in the database. If a match is not achieved, data in the text file is deleted and the program restarts waiting for a speech input, otherwise, the function corresponding to the detected command is called in order to interpret the command for the robot. And, the robot path (waypoints) are changed so that the robot follows the received command. The new waypoints are inputted to the SLAM layer, causing the control input signals to be changed. Thus, the robot starts following the command while it is performing SLAM. At the end, the data in the text file is deleted. In this way, always the last command is the one which will be followed even if the previous one is not performed completely.

In case the received command is "COME" or "FOLLOW ME", if a human is visible in the direction of the sound source in the field of view of the camera, the distance to the human is obtained from laser data, and the next control inputs (robot's new path) will be changed. So, the robot moves towards the sound source and stops at a distance of 50 centimeters from it if the command is "COME", or tracks the detected human if the detected command is "FOLLOW ME". If a specific person is going to command the robot, an adaptation of the method described in [31] is going to be applied. In this case, the system is trained to be able to detect a specific person by different gestures. The human body is considered to be composed of three parts: head, torso and feet.

It is expected a human body, with the specifications close to the trained ones be detected in the direction obtained from sound source localization. In order to minimize the computational complexity, only rectangles in which the detected sound source falls in the head part of the body model are considered as areas of interest. To each possible rectangle is assigned a constant value which is a function of the maximum likelihood of the three parts of that rectangle with the pre-trained model parts. If the constant value for a rectangle is greater than a threshold, that rectangle is assumed to contain a match to the speaker. And, the centroid of that rectangle is considered to be the center of mass of the speaker.

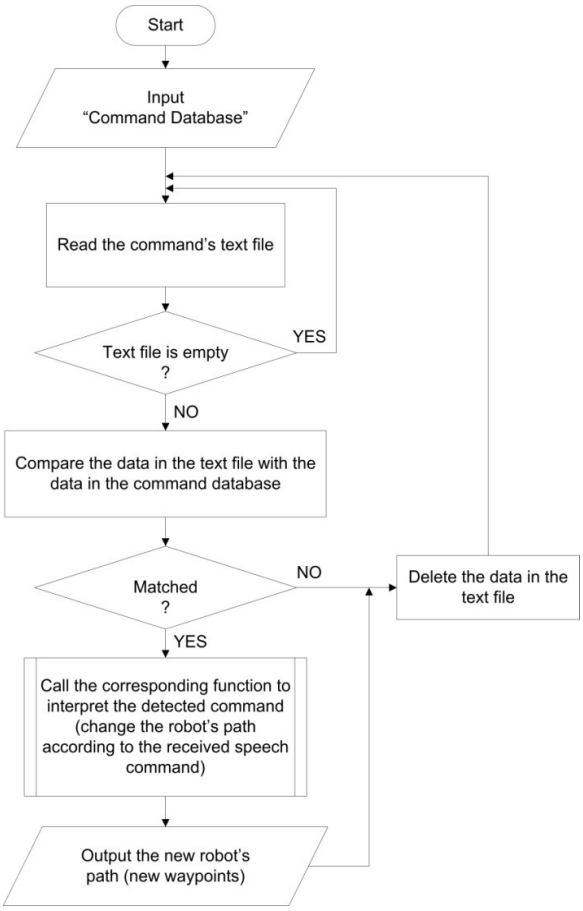


Fig. 6. Proposed speech command-based path planning

Figure 7 [31] shows the decomposition of the main rectangle into three parts based on body contour model considering the pre-trained model. To each part is assigned a value that shows the likelihood of the part to the corresponding part of the trained model. The weighted sum of the values of the three parts is the constant value assigned to the rectangle containing the detected human model (Eq. 34). Mobile parts of the body like hands and legs are excluded from the model.

$$K = 0.3K_1 + 0.6K_2 + 0.1K_3 \tag{34}$$

Once the human is detected, the robot moves towards him or follows him (depending on the received command).

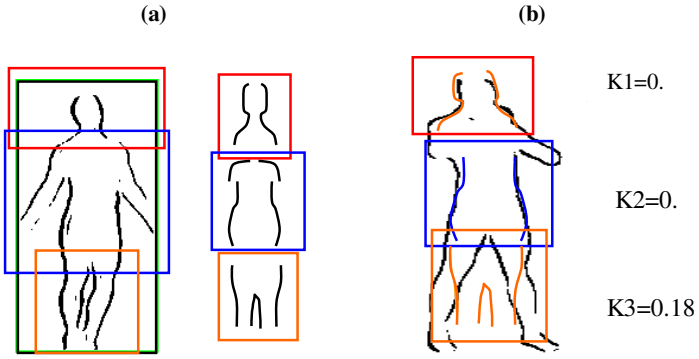


Fig. 7. a) Decomposing the rectangle around the body contour model into three parts based on the trained model (the model on the right side). b) Assigning likelihood coefficients to each part

4.3 Audio sensor model, sound source localization and speech recognition

The sound source localization method employed for audio-visual motion control of mobile robot is based on Time Delay Of Arrival (TDOA) estimation and is presented in Fig. 8. The lag time between the reception of the signals by each of the microphones is obtained by finding the maximum of the cross-correlation of the two signals received by the corresponding microphones. Then, considering that the relative geometrical position of the microphones is known, the sound source can be located.

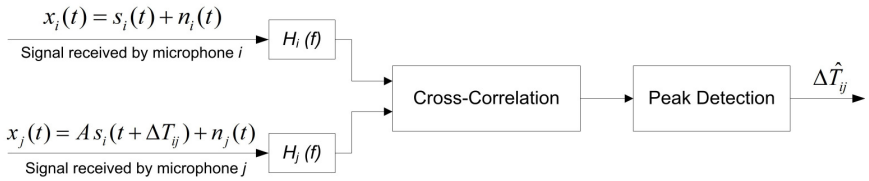


Fig. 8. Time delay of arrival estimation

Direction localization using microphone array. Assuming far field conditions, Time Delay Of Arrival (TDOA) approach is used to localize the direction of the sound source. The lag time between each pair of microphone is obtained by finding the peak in the cross correlation of the signals received by them [21]. With n microphones, there are n-1 independent cross correlations. Therefore, by finding the lag time between the first microphone and all other microphones, all lag times can be calculated:

$$\Delta T_{ij} = \Delta T_{1j} - \Delta T_{1i} \tag{35}$$

ΔT_{1j} (for j=2,3,...,n) is calculated using the following equation:

$$\Delta T_{1j} = \arg\max_{\tau} R_{1j}(\tau) \tag{36}$$

where, R_{1j} is the cross correlation between the signals received at the first microphone and the rest of the microphones, assuming that the microphones are not all positioned in the same plane (for the stability of the system of equations mentioned in Eq. 40). Since the objective of the sound source localization in the proposed system is to localize the source of the speech command and considering that the voice signal is generally low-pass, the peaks of the cross-correlations can be very wide. This problem is solved by normalizing (whitening) the spectrum of the signals prior to computing the cross-correlation [32,33]. Also, in order to increase the robustness of the signal to noise, more weight is given to the the regions in the spectrum with higher signal-to-noise ratio (SNR). Assuming that X(k) is the mean power spectral density for all the microphones at a given time and that Xn(k) is a noise estimate based on the time average of previous X(k). The noise masking weight is:

$$w(k) = \max(0.1, \frac{X(k) - \alpha X_n(k)}{X(k)}) \; ; \; \alpha < 1 \tag{37}$$

In tonal regions of the spectrum, the SNR is very high. Thus, the contribution of the signal in tonal regions is increased using the following weight function:

$$w(k) = \begin{cases} w_1(k) & if \;\; X(k) \le X_n(k) \\ w_1(k)\left(\frac{X(k)}{X_n(k)}\right)^{\gamma} & if \;\; X(k) > X_n(k) \end{cases} \quad ; \quad 0 < \gamma < 1 \tag{38}$$

Therefore, the resulting weighted cross-correlation. is:

$$R_{mj}(\tau) = \sum_{k=1}^{n-1} \frac{w^2(k) X_m(k) X_j^*(k)}{\left|X_m(k)\right|\left|X_j(k)\right|} e^{i2\pi k\tau/n} \tag{39}$$

After obtaining the lag times between all microphones, sound source localization is achieved based on the relative geometrical positions of microphones. As is illustrated in Fig. 9, let $\vec{\mathbf{s}} = (u, v, w)^T$ represent the sound source direction, for n microphones, each with location (x_i, y_i, z_i), we have:

$$\begin{bmatrix} (x_2 - x_1) & (y_2 - y_1) & (z_2 - z_1) \\ (x_3 - x_1) & (y_3 - y_1) & (z_3 - z_1) \\ \vdots & \vdots & \vdots \\ (x_n - x_1) & (y_n - y_1) & (z_n - z_1) \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} v_s.\Delta T_{12} \\ v_s.\Delta T_{13} \\ \vdots \\ v_s.\Delta T_{1n} \end{bmatrix} \quad (40)$$

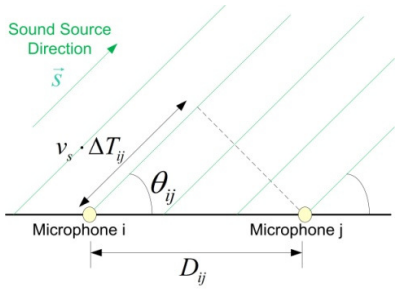


Fig. 9. Sound source localization

5 Algorithms for quality estimation of perceived speech and image information from the robot to increase the mobile robot audio visual perception

The importance of quality estimation of perceived from the robot speech and image information can be motivated as one of the possible effective methods to increase the precision of mobile robot motion control[2].

5.1 Algorithm of quality estimation of perceived speech information from the robot

There are a lot of methods [34], [35], [36], [37] and standards [38], [39], [40] for audio quality estimations. Most of them are based on subjective tests, other are objective methods and algorithms trying to obtain the precision of subjective methods. For the mobile robot speech perception quality estimation only the objective methods and algorithms can be applied.

[2] This section is based on parts of the researches done towards the PhD thesis "Development of Methods and Algorithms of Audio and Video Quality Estimation and increasing in multimedia communication systems", conducted at The French Language Faculty of Electrical Engineering, Technical University of Sofia, Bulgaria.

An important condition to choose an adequate objective algorithm is the closeness to the precision of the subjective methods widespread for speech quality estimation in communication systems [41], [42], [43], [44]. The algorithm corresponding to this condition is presented in Fig.10 and is based on the proposed method of objective speech quality estimation replacing person as estimator in subjective methods with text to speech and speech to text methods as criterion in speech quality estimation [45], [46]. The main advantages of this proposition are the elimination of the person subjective factor in speech quality estimation process and the approach of the precision of objective speech quality methods to the higher precision of subjective methods. In the Fig. 11 is presented the corresponding simulation model of the mobile robot speech perception quality estimation.

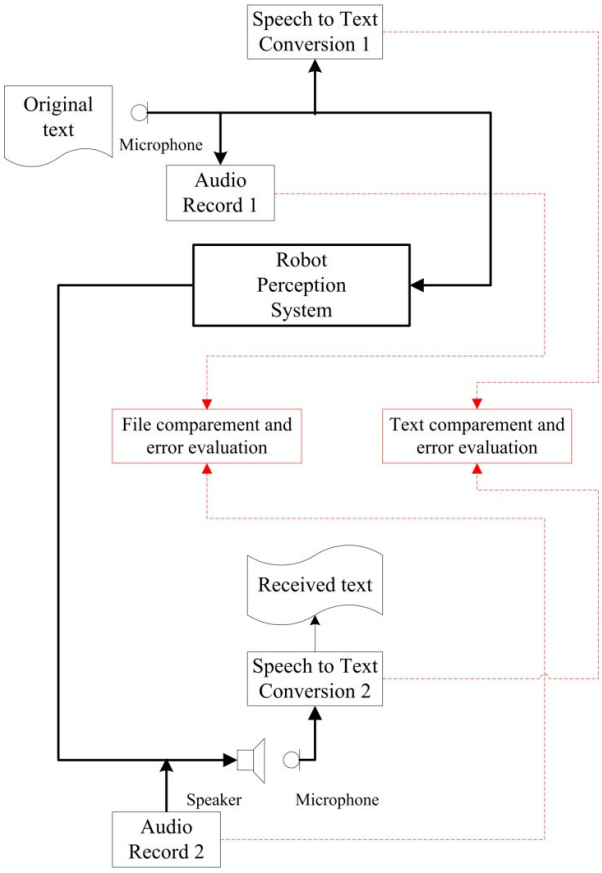


Fig. 10. Algorithm of objective quality estimation of speech robot perception

In the beginning of the algorithm is used an original text (marked as block “Original text”) from printed document, which is converted into a speech signal from a microphone connected to the mobile robot perception system (marked as block “Robot perception system”). The input speech signal is recorded as audio file (marked as block “Audio Record 1”) in the mobile robot computer perception system and simultaneously is converted into a digital text file (referred as block “Speech to Text Conversion 1”). The converted speech signal into a digital text file must be interpreted from the mobile robot as a speech command from a person speaking to the robot. In the same time the perceived from the mobile robot speech signal is reproduced by loudspeaker device (presented as “Speaker” in Fig.10) and is recorded on the computer as audio file (marked as block “Audio Record 2”). In front of and nearby the loudspeaker device is placed another microphone, which receives the speech signal for conversion into a new text file (referred as block “Speech to Text 2”).

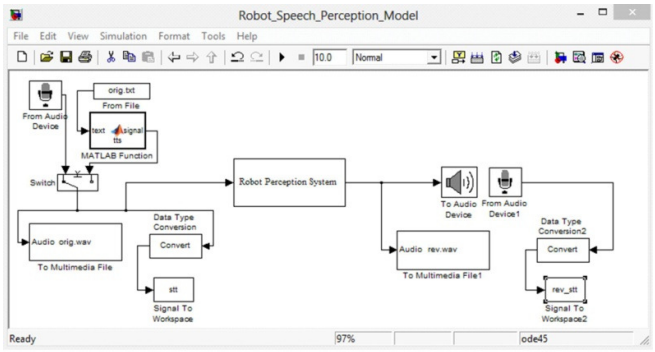


Fig. 11. Simulation model of audio input part with text to speech and audio output part with speech to text

In the simulation model on Fig. 11 are presented two types of possibilities to choose the source of the speech signal:

- real speech signal perceived direct from mobile robot microphone (From Audio Device);
- speech signal converted from a speech to text system (Data Type Conversion).

The simulation model from Fig.11 execute algorithm presented on Fig.10 in situations, when the mobile robot receive person speech commands. The quality estimation of speech command perception from mobile robot is prepared as the comparison (marked as block “Text compartment and error evaluation” in Fig.10.) and calculation of the number of incorrect received words between the two text files (Fig.11):

- text document created after direct speech to text converssion of spoken from person words as speech comands to the robot and saved as text file **stt.txt**;
- text document created after speech to text converssion of perceved from the robot words as speech comands and saved as text file **rev_stt.txt**.

As a result of comparison the error evaluations are used to define two types of quality assessment for mobile robot objective speech perception as speech command from a person:

$$OSQE_D = DNErW = NErW_{person} - NErW_{robot} \tag{41}$$

or

$$OSQE_R = RNErW = \frac{NErW_{person}}{NErW_{robot}} \tag{42}$$

where
$OSQE_D$ *and* $OSQE_R$ are the objective quality estimations of mobile robot speech perception defined as difference *(DNErW)* or as ratio *(RNErW)* between the number of erroneous words ($NErW_{robor}$) after speech to text converssion of perceved from the robot words as speech comand and the number of erroneous words *(*$NErW_{person}$*)* after direct speech to text converssion of spoken from person words as speech comand to the robot.

An additional to the proposed above objective quality estimations of mobile robot speech perception is the possibility to prepare an extra or additional estimation function for more precise objective speech quality assessment of the method and algorithm presented in Fig.10. This additional estimation is proposed to prepared as a possible comparison (marked as block "File comparement and error evaluation" in Fig.10.) between the two audio records "Audio Record 1" and "Audio Record 2":

- original speech signal saved as speech file **orig.wav**,
- received speech signal saved as speech file **rev.wav**.

6 Experimental results and discussions

Experimental results in this section are obtained by the robot Surveyor SRV-1 [47], which is equipped with a platform of sensors consisting of a Blackfin camera [48] and a Hokuyo URG-04LX-UG01 scanning laser rangefinder [49] (Fig. 12). The camera uses the OV9655 CMOS sensor [50]. The Signal-to-Noise Ratio (SNR) of the sensor and its dynamic range (ratio between the maximum and minimum measurable light intensities) are 42 dB and 50 dB, respectively. The proposed control system for audio-visual mobile robot motion control is assumed to be applied in structured indoor environments. The main system is demonstrated by simulations in Matlab, which are based on real sensor data. In parallel with the main algorithm, Microsoft Speech API is used by a program, written in visual basic, for writing the received speech signals in a text file. The speech commands are detected in the main program by reading this text file and comparing it with the command database. The commands are followed by employing SLAM [21].

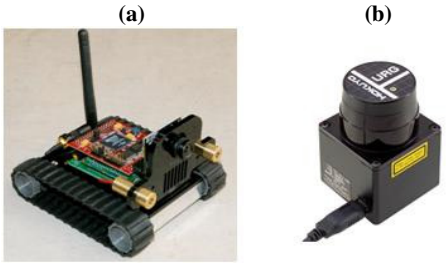


Fig. 12. a) Robot Surveyor SRV-1 equipped with a Blackfin Camera. b) Hokuyo URG-04LX-UG01 Laser Range Finder

The control noise, $\mathbf{Q}$ and the measurement noise, $\mathbf{R}$ are assumed to be:

$$\mathbf{Q} = \begin{bmatrix} (0.2)^2 & 0 \\ 0 & (1^\circ)^2 \end{bmatrix}; \qquad \mathbf{R} = \begin{bmatrix} (0.08)^2 & 0 \\ 0 & (1^\circ)^2 \end{bmatrix} \tag{43}$$

The environment of the experiments is the telecommunication laboratory (No. 1258). Measurements provide range-bearing information about the landmarks (corners) of the environment. In Simulations, the frequency of control updates is 40 Hz and observations are obtained with a frequency of 5 Hz, and the robot speed and wheelbase are assumed to be 0.1 m/s and 10 cm, respectively.

6.1 Sensor calibration

In the proposed system for audio-visual motion control of mobile robots, the camera and laser range finder must be calibrated extrinsically so that their relative position could be computed. This is useful for modeling the environment by corners, detected from laser data and associated with their corresponding vertical edges in the field of view of the camera.

Intrinsic parameters of the camera are obtained by geometric camera calibration. These parameters are used for compensation of lens distortions. The camera is calibrated practically using Bouguet's camera calibration toolbox [51]. A sequence of images of the chessboard pattern, captured by the camera from different positions with varying depth and angle, used for camera calibration, is shown in Fig. 13.

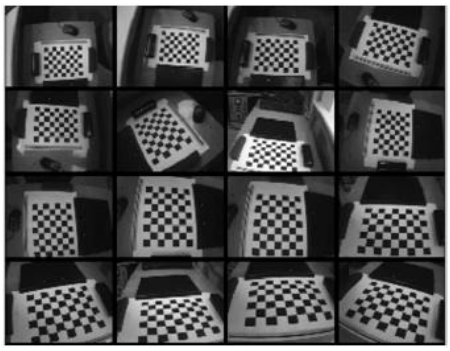


Fig. 13. The sequence of images of the test pattern used for geometric camera calibration

After selecting the extreme grid corners for each image and providing geometrical information about the dimensions of the grid cells ($30^{mm} \times 30^{mm}$), corner extraction is performed and camera calibration parameters are computed by minimizing the least square error between the observed corner coordinates and the coordinates computed based on the calibration model. Figure 14 demonstrates the extracted corners from two of the images.

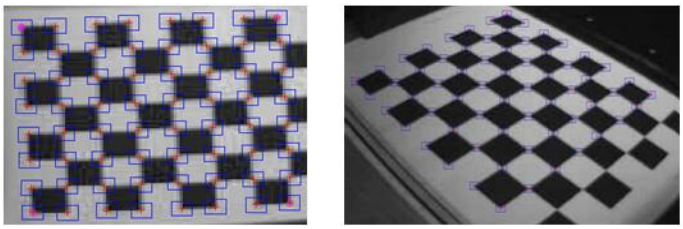


Fig. 14. Corner extraction performed in camera calibration

The intrinsic parameters of the camera are presented in Table 3. They are:

$$\left\{(f_u, f_v), (u_0, v_0), (k_1^{(r)}, k_2^{(r)}), (k_1^{(t)}, k_2^{(t)})\right\}; \tag{44}$$

where

- (f_u, f_v) is the focal length in pixels (it includes the scaling factors, too);
- u_0, v_0 are the coordinates of the principal point in pixels;
- $k_1^{(r)}, k_2^{(r)}$ present radial lens distortion coefficients and are related to the radial distortion coefficients by $k_1^{(r)} = f^3 k_1$ and $k_2^{(r)} = f^5 k_2$.

– $k_1^{(t)}, k_2^{(t)}$ present tangential lens distortion coefficients and are related to the tangential distortion coefficients by $k_1^{(t)} = f^2 p_1$ and $k_2^{(t)} = f^2 p_2$.

Table 3. Results of geometric calibration of the camera

Parameter	Value	standard dev.
f_u (pixels)	301.49221	1.71848
f_v(pixels)	303.21885	1.57392
u_0(pixels)	155.45581	1.18310
v_0(pixels)	139.29446	1.82011
$k_1^{(r)}$	-0.43039	0.00763
$k_2^{(r)}$	0.24063	0.01870
$k_1^{(t)}$	-0.00382	0.00195
$k_2^{(t)}$	-0.00297	0.00052

It is visible in Table 3 that lens tangential distortion is negligible. On the other hand, there is a considerable radial lens distortion, and image correction is performed on the images captured by the camera based on radial lens distortion coefficients.

Another useful information provided by the geometric camera calibration is that the relative position of each of the images of the sequence with regard to the local frame of the camera is obtained as extrinsic parameters of each image (Fig. 15).

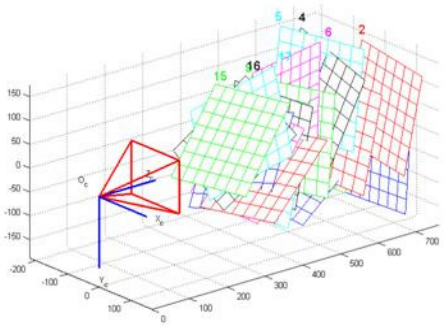


Fig. 15. Extrinsic parameters (camera-centered)

Therefore, considering that the position of all corners is now known with respect to the camera frame, by finding the board corners in laser scan data for each chessboard

position, laser rangefinder's relative position with respect to the camera is iteratively calculated minimizing the squared error.

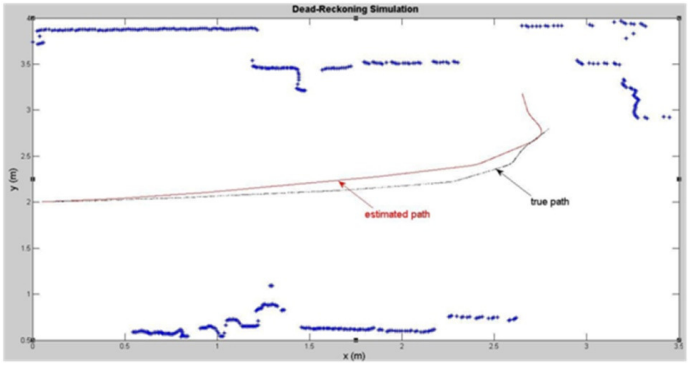
After calibration of the sensors, their systematic errors are compensated and measurements can be modeled by Gaussian distribution containing uncertainty caused by random errors. Also, assuming measurements to have Gaussian distribution in SLAM generates results with good precision. Additionally, because of extrinsic calibration of laser rangefinder and the camera, correspondence is found between vertical edges in the image received from the camera and corners in laser data.

6.2 Robot navigation based on EKF-SLAM

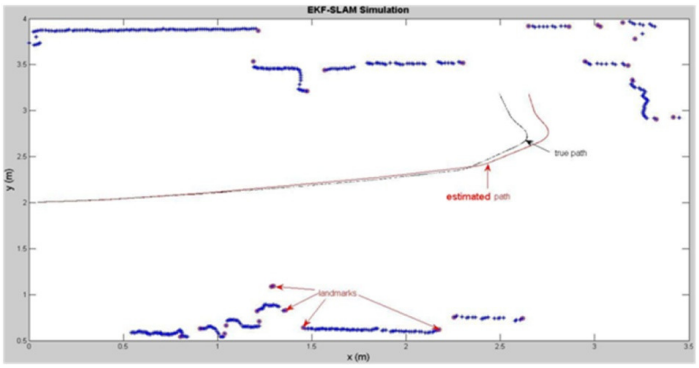
In the following simulations, the accuracy of robot navigation based on EKF-SLAM is compared to robot navigation based on dead-reckoning. Therefore, it is assumed that the same path is planned within the same environment, and that the robot starts from the same point in all the experiments. Each of the algorithms are applied 40 times under the same conditions, and the result with the highest average localization error (the worst result) among the 40 experiments represents each of the algorithms. Figure 16 [21] demonstrates simulation results and robot localization absolute and average errors for each of the algorithms.

It is obvious that in mobile robot positioning based on dead-reckoning, the accuracy decreases over time. Figure 16.c shows that positioning error is accumulated over time because random errors of proprioceptive sensor measurements accumulate and lead to incremental uncertainties in the robot position estimation over time. In about 3.2 meters of robot navigation based on dead-reckoning, the average error relative to the traveled path is more than 6%. SLAM algorithms are applied to correct the high localization error in dead-reckoning based on environmental perceptions acquired from exteroceptive sensors. In EKF-SLAM, the error is decreased to a great extent. The average error from start point until the robot reaches the goal point is reduced from 0.2 meters in dead-reckoning to 0.08 meters in EKF-SLAM. In robot navigation based on EKF-SLAM, after 3.2 meters of robot displacement, the average error relative to the traveled path is more than 2.5%.

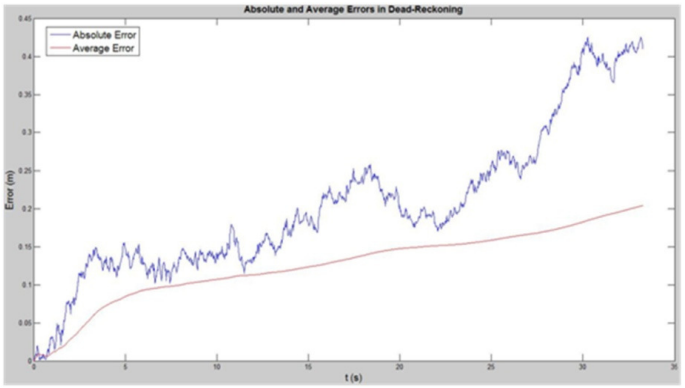
(a)



(b)



(c)



(d)

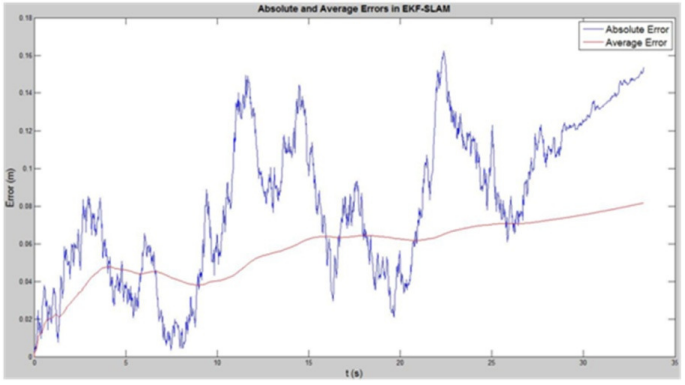


Fig. 16. a) Dead-Reckoning simulation; b) EKF-SLAM simulation; c) Absolute and average errors in dead-reckoning; d) Absolute and average errors in EKF-SLAM

6.3 Experimental results from simulations of the proposed objective speech quality estimation based on original and received texts comparison

After running the simulation model presented in Fig. 11 are calculated the defined two types of quality assessment for mobile robot objective speech perception as speech command from a person, using equations (41 and 42). Some of the important results from the simulations are shown in Fig.17, where are presented the results from a simple example of one of the simulations:

- part of the text document (file **stt.txt**) after direct speech to text conversion of spoken from person words as speech comands to the robot;
- part text document (file **rev_stt.txt**) after speech to text conversion of perceved from the robot words as speech comands.
- It can be seen from Fig. 17, that there are differences between the number of erroneous words (marked in yellow color) as speech comands to the robot in the text documents **stt.txt** and **rev_stt.txt**. With This difference is used to calculate with equations (41 and 42), the values the objective speech quality estimation for the robot speech perception.

file **stt.txt**
start; come; follow me; stop moving.

start; go to; turn left; stop moving.

start; follow me; turn right; turn left; stop moving.

start; go to; turn right; turn left; stop moving.

start; go to; stop at; start; turn right; go back; stop moving.

start; come; turn left; follow me; stop moving.

start; go to; stop at; start; turn left; turn left; stop moving.

start; follow me; stop at; start; turn right; stop moving.

file **rev_stt.txt**
start; come; follow me; stop moving.

start; go to; turn left; stop moving.

start; follow me; turn right; turn left; stop moving.

start; go to; turn right; turn left; stop moving.

start; go to; stop at; start; turn right; go back; stop moving.

start; come; turn left; follow me; stop moving.

start; go to; stop at; start; turn left; turn left; stop moving.

start; follow me; stop at; start; turn right; stop moving.

Fig. 17. Parts of the text documents stt.txt rev_stt.txt after speech to text in transformation and receiving parts with erroneous spoken words marked in yellow color

– For the example in Fig.17 the concrete values of $NErW_{robot}$ and $NErW_{person}$ are: $NErW_{robot} = 12$ and $NErW_{person} = 6$ Then from the equations (41 and 42) are calculated the following values of the objective speech quality estimation for the robot speech perception:

$$OSQE_D = DNErW = NErW_{person} - NErW_{robot} = 6 - 12 = -6 \tag{45}$$

$$OSQE_R = RNErW = \frac{NErW_{person}}{NErW_{robot}} = \frac{6}{12} = 0.5 \tag{46}$$

Therefore, from the equations (45 and 46) can be concluded that calculated values of two types of the objective speech quality estimation gives a quantitative objective notion for the robot speech perception useful for estimation the precision of mobile robot motion control and guidance with speech commands from a person.

7 Conclusion

The proposed audio-visual perception system in this chapter is used for joined audio visual mobile robot motion control employing audio-visual and range information perceived from robot sensors (microphone array, video camera, and laser rangefinder). The algorithms developed for robot motion control through speech commands and robot navigation by performing EKF-SLAM, that assumes vertical edges as the environment landmarks, are based on perceived audio information from microphone and visual and range information of camera images and 2D laser rangefinder, respec-

tively. The way of modeling the environment as vertical edges in the camera image associated to corners in range information from the laser rangefinder has the advantage that because there is not high feature clutter, the problem of quadratic complexity of the EKF-SLAM is solved to a great extent. The importance of mobile robot audio visual perception to achieve a defined motion control precision is estimated with the proposed in this chapter algorithm of objective quality estimation of speech robot perception. It is show that the application of well known Microsoft Speech to Text and inverse Text to Speech algorithms allow to replace person as estimator of speech quality and to approach the precision of objective speech quality estimation methods of mobile robot audio perception to corresponding subjective methods. It is necessary to mentioned that all the presented in this chapter results are subject of researches done towards the two PhD thesis's: "Development of Methods and Algorithms for Audio-Visual Mobile Robot Motion Control" and "Development of Methods and Algorithms of Audio and Video Quality Estimation and increasing in multimedia communication systems", conducted at The French Language Faculty of Electrical Engineering, Technical University of Sofia, Bulgaria.

References

1. Intelligent Service Robotics. Editor-in-Chief: Il H. Suh, Springer, ISSN: 1861-2776, Journal No11370
2. Jarvis R.. Intelligent Robotics: Past, Present and Future. International Journal of Computer Science and Applications, ©Technomathematics Research Foundation, Vol.5, No.3, 2008, pp.23-35
3. Bittermann M. S., I. Sevil Sariyildiz and Özer Ciftcioglu. Visual perception in design and robotics. Journal of Integrated Computer-Aided Engineering-Informatics in Control, Automation and Robotics, Vol.14 Issue 1, January 2007, pp.73-91
4. Bigun J., Vision with direction, Springer Verlag, 2006
5. Adams B., C. Breazeal, R.A. Brooks, B. Scassellati, Humanoid robots: a new kind of tool, Intelligent Systems and Their Applications, IEEE Vol.15, No. 4 (2000), 25-31.
6. Eckmiller R., O. Baruth and D. Neumann, On human factors for interactive man-machine vision: requirements of the neural visual system to transform objects into percepts, Proc. IEEE World Congress on Computational Intelligence WCCI 2006 - Int. Joint Conf. on Neural Networks, Vancouver, Canada, July 16-21, 2006, pp.99-703.
7. Demmel J., G. Lafferriere, J. Schwartz, and M. Sharir. Theoretical and experimental studies using a multifinger planar manipulator. In IEEE International Conference on Robotics and Automation, 1988, pp.390–395
8. Murray R. M., S. S. Sastry. Grasping and manipulation using multi fingered robot hands. In R. W. Brockett, editor, Robotics: Proceedings of Symposia in Applied Mathematics, Vol.41, American Mathematical Society, 1990, pp.91–128
9. Manocha D. and J. F. Canny. Real time inverse kinematics for general 6R manipulators. Technical Report ESRC 92-2, University of California, Berkeley, 1992
10. Jarvis R., Ho, N. and Byrne. J.B, Autonomous Robot navigation in Cyber and Real Worlds, CyberWorlds 2007, Hanover, Germany, Oct. 24th to 27th, 2007, pp. 66-73
11. Siegwart R., Illah R. Nourbakhsh. Introduction to Autonomous Mobile Robots. A Bradford Book, The MIT Press Cambridge, Massachusetts, London, England 2004

12. Adams Mm. Sensor Modelling, Design and Data Processing for Autonomous Navigation. World Scientific Series in Robotics and Intelligent Systems. Singapore, World Scientific Publishing Co. Ltd., 1999
13. Borenstein, J., H.R. Everet, L.Feng. Navigating Mobile Robots, Systems and Techniques. Natick, MA, A.K. Peters, Ltd., 1996
14. Humanoid Robots. New Developments. Edited by Armando Carlos de Pina Filho. Published by Advanced Robotic Systems International and I-Tech, Vienna Austria, 2010.
15. Kuffner J.J., K.Nishiwaki, S.Kagami, M.Inaba.. Motion Planning for Humanoid Robots under Obstacle and Dynamic Balance Constraints, Proceedings of IEEE International Conference on Robotics and Automation, 2001, pp.
16. Favorskaya M. Recognition of dynamic visual images based on group transformations. Journal Pattern Recognition and Image Analysis, Springer-Verlag New York, Inc. Secaucus, NJ, USA ,Vol. 22, Issue 1, March 2012, pp. 180-187
17. Favorskaya M., Dm. Pyankov, Al. Popov. Motion Estimations based on Ivariant Moments for Frames Interpolation in Stereovision. 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013, Elsevier, Procedia Computer Science 22 (2013), pp.1102 – 1111.
18. Wit C. Trends in mobile robot and vehicle control, in Control Problems in Robotics, B. Siciliano and K. P. Valavanis (eds.), London, U.K., Springer-Verlag, 1998
19. Bekiarski Al., Sn. Pleshkova. Microphone array beamforming for mobile robot. Proceeding CSECS'09 Proceedings of the 8th WSEAS International Conference on Circuits, systems, electronics, control & signal processing, pp. 146-149
20. Favorskaya M., Нахождание движущихся видеообектов с приминением локальных 3D-структурных тензоров. Вестник Сибирского государственного аэрокосмического университета им. академика М.Ф. Решетнева. 2009. № 2. С.141-146
21. Dehkharghani, Sh. Sehati. Development of Methods and Algorithms for Audio-Visual Mobile Robot Motion Control (Doctoral dissertation), 2013
22. Dehkharghani, Sh. Sehati, & Pleshkova, S. "Geometric Thermal Infrared Camera Calibration For Target Tracking by a Mobile Robot." Comptes rendus de l'Academie bulgare des Sciences, vol. 67(2), 2014, accepted for publication.
23. Heikkila, J., & Olli, S. A four-step camera calibration procedure with implicit image correction. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997, pp. 1106-1112
24. Abdel-Aziz, Y. I., & Karara, H. M.. Direct linear transformation into object space coordinates in close-range photogrammetry. Proceedings of Symposium on Close-Range Photogrammetry, Urbana, Illinois, , 1971, pp. 1-18.
25. Levenberg, K., A Method for the Solution of Certain Nonlinear Problems in Least Squares, Quarterly of Applied Mathematics, Vol. 2, No. 2, , 1994, pp. 164-168
26. Benjamin, J. R. , & Cornell, C. A.. Probability, Statistics, & Decision for Civil Engineers. New York: McGraw-Hill, 1970, pp. 9-27
27. Dehkharghani, Sh. Sehati, Al. Bekiarski, Sn. Pleshkova. Application of Probabilistic Methods in Mobile Robots Audio Visual Motion Control Combined with Laser Range Finder Distance Measurements. Advances in Circuits, Systems, Automation and Mechanics, December 2012, pp. 91-98
28. Dehkharghani, Sh. Sehati, Al. Bekiarski, Sn. Pleshkova. Method and Algorithm for Precise Estimation of Joined Audio Visual Robot Control. Iran's Third International Conference on Industrial Automation, Tehran, 22-23.01.2013, pp.
29. Hough, P. VC.. Method and means for recognizing complex patterns. U.S. Patent No. 3,069,654. 18.12.1962.

30. Canny, J. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 6,1986, pp. 679-698.
31. Venkov, P., Al. Bekiarski, Sh. Dehkharghani Sehati, Sn Pleshkova. Search and tracking of targets with mobile robot by using audio-visual information. Proceedings of the International Conference on Automation and Informatics (CAI'10), Sofia, 2010, pp.463-469
32. Omologo, M., & Svaizer, P. Acoustic event localization using a crosspower-spectrum phase based technique. Proceedings of IEEE International Conference on Acoustics, Speech, & Signal processing, vol. 2, 1994, pp. 273-276
33. Valin, J-M., François Michaud, Jean Rouat, & Dominic Létourneau. Robust sound source localization using a microphone array on a mobile robot. Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003), vol. 2, 2003.pp. 1228-1233
34. Ping Ji, L. Benyuan, D. Towsley, J. Kurose, "Modeling Frame-level Errors in GSM Wireless Channels", IEEE Globecom, Internet Performance Symposium, 2002
35. Mohamed S., G. Rubino, M. Varela. "A method for quantitative evolution of audio quality over packet networks and its comparison with existing techniques", in Measurement of Speech and Audio Quality in Networks (MESAQIN), 2004
36. Hu Y., P. Loizou. "Subjective comparison of speech enhancement algorithms," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, vol. 1, 2006, pp. 153–156
37. Kondo K. "Subjective Quality Measurement of Speech. Its Evaluation, Estimation and Applications". Springer, Signals and Communication Technology, 2012
38. ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs", International Telecommunication Union, Geneva, Switzerland, February, 2001
39. ITU-T, ITU-T Rec. P. 835 "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm", 2003
40. L. Malfait, J. Berger, and M. Kastner, "P.563-the ITU-T standard for single-ended speech quality assessment," IEEE Trans. Audio, Speech,Lang. Process., vol. 14, no. 6, pp. 1924–1934, Nov. 2006
41. ITU-T Recommendation P.800: Methods for subjective determination of transmission quality, Aug. 1996
42. ITU-R Rec. BS.1116, "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunication Union, Geneva, Switzerland, March, 1994
43. Thorpe L., "Subjective evaluation of speech compression codes and other non-linear voice-path devices for telephony applications," Int. J. Speech Technol., vol. 2, 1999, pp. 273-288
44. Pleshkova Sn., K. Peeva. Simulation of Different Types of Voice Communication Systems Used for Speech Quality Estimation with Applying Speech to Text as Objective Criterion of Audio Quality, International Journal of Emerging Technologies in Computational and Applied Sciences, Issue 6 Vol. 2, 2013, pp. 107-111
45. Pleshkova Sn, K. Peeva. Application of Speech to Text as Criterion of Audio Quality Estimation in Multimedia Communication Systems, Sofia, CEMA'2013, pp.92-95
46. Pleshkova-Bekjarska Sn. Simulation Analysis of Speech Quality Dependence from Communication Channel Type and Channel Coding Methods, International Journal of Emerging Technologies in Computational and Applied Sciences, Issue 6, Vol. 1, 2013, pp. 8-12
47. Surveyor SRV-1 Blackfin Robot Surveyor Corporation. http://www.surveyor.com/SRV_info.html

48. Surveyor SRV-1 Blackfin Camera Surveyor Corporation.
http://www.surveyor.com/blackfin/
49. Scanning range finder (SOKUIKI sensor): URG-04LX-UG01 Hokuyo Corporation.
http://www.hokuyo-aut.jp/02sensor/07scanner/urg_04lx_ug01.html
50. Datasheet available in: http://www.surveyor.com/blackfin/OV9655-datasheet.pdf
51. Bouguet J. Camera Calibration Toolbox for Matlab
http://www.vision.caltech.edu/bouguetj/calib_doc/