Human Tracking Using Delphi ESR-Vision Fusion in Complex Environments

T. Wang, R. Aggarwal, and A. K. Somani
Electrical and Computer Engineering, Iowa State University, Ames, Iowa, USA, 50010
{tengw, rka, arun} @iastate.edu

Abstract—A variety of UGV (Unmanned Ground Vehicle) applications pose the challenge that UGVs need to handle human detection and tracking in complex environments that include dusty, smoky and foggy conditions. These environments make a vision-based human tracking-by-detection system ineffective. To cope with this challenge, we build a radar-vision fusion system, utilizing a 77GHz 2D Delphi ESR (Electronically Scanning Radar) and a CCD camera. Our fusion system utilizes radar returns to generate ROIs (Regions of Interests) and then employs a vision-based human detection technique to validate ROIs. Considering that human are weak targets for a 77GHz radar sensor due to their smaller sizes and weaker reflectivity, we develop a human tracking approach to recover from intermittent human misses. This improves the accuracy of our multi-sensor system. We design experiments to study the behavior of Delphi ESR for human detection. We also characterize Delphi ESR’s measurement noise. Using the derived Gaussian noise model parameters, we develop a novel human tracking approach using Kalman filter. We also describe, in detail, an approach to map radar returns to image plane for generating ROIs. A set of real-world experiments show the effectiveness of our approach in human tracking and radar-vision registration.

Keywords: Radar-Vision System, Delphi ESR, Human Tracking, Kalman Filter, ROI Generation

1. Introduction

An UGV is defined as a vehicle that operates on ground without an onboard human operator presence. Currently, UGVs are being widely deployed for many smart video surveillance systems and a variety of other applications including farming and mining. For such engineering applications of UGVs, real-time and reliable human detection and tracking in their surroundings is required in order to operate them safely and securely.

Human detection and tracking with a monocular camera has been studied widely. In recent years, HOG (Histogram of Oriented Gradients) descriptor [1] and DPMs (Deformable Part-based Models) [2] are widely used in human detection applications, and show excellent performance in static images. Shu et al. [3] and Kim et al. [4] explored DPMs and HOG descriptors to multiple people tracking in crowded scene, respectively: Shu et al. proposed a robust part-based framework to handle partial occlusion and Kim et al. developed a MAP-based online data association approach to address new detection and missing detection problems.

Performances of these vision-based human tracking-by-detection systems are affected by accuracy of their human detection algorithms. Main ideas of these human detection algorithms are to characterize human by pre-defined features and utilize these discriminative features to recognize human targets in images.

It is becoming common for UGV systems to work in weather conditions, such as foggy and cloudy. In addition, the UGV system is focusing more on highly unstructured complex environments, like off-road scenarios. The off-road scenarios offer dusty and smoky environments. One typical example is autonomous surface miners. Surface miners generate dusts when they are operating. Fig. 1 depicts an example of this. These dusty, smoky environments degrade image quality and destroy the discriminative features for human recognition to a great extent. This leads to a significant performance degradation of tracking-by-detection systems.

Fig. 1: Dusty environments where two human targets are visible but not detectable by vision-based human detection algorithms.

1.1 A Solution

Due to the complex nature of the working environment, UGV systems need to be equipped with many different types of sensors to deal with the issues of environmental perception and recognition. A millimeter wave (MMW) radar is an attractive choice due to the fact that it possess all-weather operating capability and can thus penetrate fog, rain, and dust. In recent years, commercial 2D MMW radars have become more available and affordable due to their adoption in automotive applications. However, a 2D radar sensor is
limited to providing the information on radar returns which can be used for detecting ROIs. It does not have the ability to provide object discrimination.

Given the challenges of vision or radar based human tracking systems alone, we develop a radar-vision fusion system to deal with human detection and tracking in dusty, foggy and similar complex environments. Our system deploys a 2D Delphi ESR and a CCD camera, with the CCD camera put directly above the Delphi ESR to maximize their overlapping FOV (Field of View) as shown in Fig. 2. During operation, our radar-vision fusion system is synchronized to receive image and radar data in real time. For each data frame, our system utilizes radar returns to generate ROIs and employs additional local contrast enhancement techniques to enhance features in ROIs. Next, we employ a pre-trained classifier with tuning parameters to validate ROIs. It is worth mentioning that our system allows for a weaker classifier for human detection compared to a vision-based human tracking-by-detection system, as radar data processing helps reduce most of the possible false positives. This is attractive as these dusty and foggy environments destroy some discriminative human features in images and weaker modules are needed to be used.

A fusion system driven by radar needs a set of very reliable radar returns, as any radar misses cannot be recovered by the vision system. Since human are weak targets for 77GHz Delphi ESR, and deliver a discrete set of responses, we develop a human tracking approach to obtain reliable radar returns from human targets.

2. Behaviors of Delphi ESR in Human Detection

There are a number of commercially available automotive radar sensors produced by several manufacturers. We chose a 77GHz Delphi ESR for human detection purpose. This is because it provides mid-range (60m) measurements with a wide field of view (+/- 45°). This characteristics allows human targets across the width of the equipped UGV to be detected. In this project, the ESR unit is programmed to run in 64-Track mode. In this mode the ESR attempts to identify and take measurements of 64 targets every 50 milliseconds. For each track, Delphi ESR provides the range R and azimuth θ information of the target.

We conduct experiments to establish how Delphi ESR performs in applications requiring human detection. The outcomes of one experiment are shown in Fig. 3. The black line in Fig. 3a represents a predefined walking trajectory of a human. During experiment, our system is positioned at a fixed location, approximately two meters away from the start of the trajectory. The human target walks at a constant speed following the depicted path.

The sequence of ESR returns from the human target is presented in Fig. 3b, where indicator = 0 or 1 represents that Delphi ESR misses or successfully detects the human target for a specific image frame, respectively. We make the following observations from the sequence of radar returns:

- The whole experiment process outputs a total of 270 data frames, and the radar missed the human target 26 times. A person is only visible to a radar system when the person scatters back enough signal for the radar transmitter. This implies that the bigger the object is, the better the chances are that the target is detected. The radar misses can be explained by the fact that humans are bordering on the size of “small” objects for a 77GHz radar system. They are not prohibitively small, but small enough to be missed sometimes.
- It is worth mentioning that most of the human misses occurred at the beginning or at the end of the experiment when the human target was close to the radar sensor. For short range, due to ESR’s narrow elevation FOV, the portion of the human body that is illuminated by the radar is small. Therefore, the effective radar cross-section is small, which makes the human target even harder to detect. In case human are slightly farther than 2 meters, the detection probability will increase.
- Human misses are intermittent, rather than continuous. For this specific experiment, we noticed that there were at most two consecutive misses. The “intermittent” property of misses suggests that the misses could be recovered with help of additional tracking techniques.
- It is possible that there may be multiple radar returns from a single human target. An example of this phenomena is shown in Fig. 4. Fig. 4a and Fig. 4b show

1.2 Paper Organization

This paper is organized as follows. Section 2 presents our experiments with Delphi ESR to understand its behavior for the human detection problem. Section 3 explains our approach to model Delphi ESR’s measurement noise. Next we describe our human tracking approach and develop the need for the use of Kalman filter. The mapping between 2D radar space and 2D image plane to generate ROIs is discussed in Section 4. Performances of our human tracking approach as well as radar-vision registration method are analyzed in Section 5. The paper is concluded in Section 6.
the image frame and their corresponding radar detection map, respectively. In the radar detection map, both radar returns in the black circle are from the human target. This phenomena can be explained by the fact that a single human in clean background glows to the radar like a lightbulb, which may be interpreted by the radar as multiple targets, each with slight offset. This represents multipath errors. As a result, we need to incorporate a clustering algorithm that can treat each radar return cluster as a single target, when we design a tracking algorithm to recover human misses.

3. Human Tracking Using Delphi ESR

Considering the fact that humans are weak targets for radar detection as well as that human misses are intermittent, we design an efficient tracking model for Delphi ESR to recover from the human misses. We choose to employ Kalman filtering to recover from the missed data. For the effective use of this approach, we design an experiment to characterize Delphi ESR’s measurement noise in X and Y directions (i.e., \( n_x \) and \( n_y \)), separately. Take \( n_y \) as an example. We compute Y-offsets of these radar returns relative to the true position point, and build a frequency histogram using the Y-offset samples. The histogram is plotted in Fig. 5a. We make an observation that the histogram is kind of symmetric with most of the frequency counts bunched in the middle bin centered at 0 and with the counts dying off out in the tails. From the observation, we conclude that it is reasonable to model \( n_y \) using normal distribution with mean \( \mu_y = 0 \). We then compute the variance \( \sigma_y^2 \) and yield that \( \sigma_y^2 = 0.0815^2 \). The fitted normal distribution \( \mathcal{N}(\mu_y, \sigma_y^2) \) is plotted.

![Fig. 3: An example of human detection with Delphi ESR](image1)

![Fig. 4: One specific example of multiple radar returns from a single human target. Radar returns in the black circle (i.e., (6.6, 4.4) and (7.0, 3.5) expressed in \((R, \theta)\) format) are both from the human target, and others are from left buildings.](image2)
in red in Fig. 6a. The same calculations are also done to build frequency histogram from X-offset samples, and model $n_x$ using normal distribution with mean $\mu_x = 0$ and variance $\sigma_x^2 = 0.1346^2$. The frequency histogram and the estimated normal distribution $N(\mu_x, \sigma_x^2)$ of $n_x$ are depicted in Fig. 6b, in blue and red respectively. Mean squared errors of estimator $N(\mu_y, \sigma_y^2)$ and $N(\mu_x, \sigma_x^2)$ are equal to 0.0527 and 0.0895, respectively.

3.2 Human Tracking with Kalman Filter

Delphi ESR can report as many as 64 pairs of range-azimuth $(R, \theta)$ observations in each scan. Each range-azimuth pair represents the location of a scattering center (SC). As mentioned in Section 2, there might be several SCs from a single human target. Therefore, we first employ K-means clustering approach to put radar returns into different target clusters (TCs), with each TC representing a single target. The number of target clusters $K$ is determined by Silhouette Coefficients method. We employ Kalman filtering technique to help radar track its TCs, given that measurement noises of Delphi ESR in X and Y direction both follow the Gaussian distribution.

The state vector of each TC at frame $k$ is defined by the following equation.

$$X_k = [x_k, y_k, vx_k, vy_k]^T$$  \hspace{1cm} (1)

In Eq. 1, $T$ stands for the transpose operation; $(x_k, y_k)$ and $(vx_k, vy_k)$ are the TC’s center location and velocity in the radar cartesian coordinate system, respectively. The constant-velocity model in Eq. 2 is used to describe the dynamics of the TC.

$$
\begin{bmatrix}
x_{k+1} \\
y_{k+1} \\
vx_{k+1} \\
vvy_{k+1}
\end{bmatrix} =
\begin{bmatrix}
1 & 0 & \Delta t & 0 \\
0 & 1 & 0 & \Delta t \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix} \times
\begin{bmatrix}
x_k \\
y_k \\
vx_k \\
vvy_k
\end{bmatrix}
$$  \hspace{1cm} (2)

In Eq. 2, $\Delta t$ is the time between consecutive frames, and is assumed to be constant. As Delphi ESR outputs the relative position of the target, the measurement state vector is defined by the following equation.

$$Z_k = [x_k, y_k]^T$$  \hspace{1cm} (3)

The measurement equation can then be described by the following equation.

$$
\begin{bmatrix}
x_k \\
y_k
\end{bmatrix} =
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0
\end{bmatrix} \times
\begin{bmatrix}
x_k \\
y_k \\
vx_k \\
vvy_k
\end{bmatrix} + n
$$  \hspace{1cm} (4)

In Eq. 4, $n = [n_x, n_y]^T \sim N(0, R)$ is two dimensional measurement noise vector, with $n_x$ and $n_y$ representing measurement noise in X and Y direction, respectively. Based on our measurement noise model, we obtain $R$ as follows.

$$R = \begin{bmatrix}
\sigma_x^2 & cov(n_x, n_y) \\
cov(n_x, n_y) & \sigma_y^2
\end{bmatrix} = \begin{bmatrix}
0.1346^2 & -0.0026 \\
-0.0026 & 0.0815^2
\end{bmatrix}
$$  \hspace{1cm} (5)

The state transition and measurement equations, as shown in Eq. 2 and Eq. 4 respectively, join together to form the Kalman filter-based tracking model of a TC. The state transition equation can also be treated as a predictor when there are no radar returns from the corresponding target object, and therefore help radar track the target object.

3.3 Multiple Human Tracking Based on MAP Association

In complex environments with multiple human targets, we need to associate current TC observations with recently updated tracking models in order to update the state information of each tracking model. To achieve this, we employ a MAP-based association approach [4]. Its main idea is to encode the problem of multiple people tracking as a node matching problem. The approach adds an extra node for the automatic tracking initialization and formulate the MAP problem to associate detection observations in current time step and tracking models from the last frame.

Before proceeding, we define the following notations to facilitate explanation.

- $tc_k^i = (cx_k^i, cy_k^i)$: $i$th TC observation from $k$th radar frame where $(cx_k^i, cy_k^i)$ is the TC’s center location.
- $v_k^j = (vx_k^j, vy_k^j, vx_k^j, vy_k^j)$: $j$th tracking model after $k$th data frame, where $(vx_k^j, vy_k^j)$ and $(vx_k^j, vy_k^j)$ are position and velocity of the tracking model, respectively.
• \( P(t_{c_k}^i, v_{k-1}^j) \): similarity score between \( i \)th current TC observation and \( j \)th existing tracking model.

For the radar tracking problem, we define the similarity score as follows to apply the MAP-based association approach.

\[
P(t_{c_k}^i, v_{k-1}^j) = \frac{1}{Z} \exp(-\text{dist}(t_{c_k}^i, (v_{k-1}^j)_{\text{pos}}) + (v_{k-1}^j)_{\text{vel}} \cdot \Delta t)
\]

(6)

In Eq. 6, \( Z \) is the normalizing term, and \( \text{dist} \) is the Euclidean distance. The subscripts \( \text{pos} \) and \( \text{vel} \) stand for the position and velocity of tracking model in Cartesian coordinates, respectively.

In addition, a score is assigned to each tracking model to evaluate its accurate status. When the tracking model is created at frame time of \( k \), its accurate score \( L(k) \) is initialized by Eq. 7.

\[
L(k) = C
\]

(7)

In Eq. 7, \( C \) is a constant. \( L(k) \) is updated as follows.

\[
L(k) = \begin{cases} 
C, & \text{if matching TC observation exists} \\
L(k-1) - 1, & \text{otherwise}
\end{cases}
\]

(8)

Once we obtain the evolution curve of the accurate score, a tracking model is deleted when the accurate score \( L(k) \) is decreased to 0. Note that \( C \) refers to the maximum number of allowable consecutive human misses. We set \( C \) to be 5 in our experiments. This is because we observed from experiments that consecutive misses for human target was smaller than 5 with probability as high as 95%.

Algorithm 1 describe, in detail, our human tracking approach using Delphi ESR based on Kalman filter. At the start of the program (i.e., \( k = 1 \)), we cluster radar returns into TCs and initiate a new tracking model for each TC using the TC’s center position information (step 1 & 2). In the following frames (i.e., \( k > 1 \)), we create TCs from raw radar returns, and associate current TC observations with existing tracking models based on MAP-based association approach (step 3 & 4). In case that Delphi ESR successfully detects an existing target in current frame, we update the state vector of the target’s tracking model based on new TC observation (case 1); otherwise we predict the new position of the target using its existing tracking model (case 2). In another case that a new target appears, we initiate a new tracking model for the target (case 3).

**Algorithm 1: Human Tracking with Kalman Filter**

1. At frame \( k = 1 \)
2. Step 1: Employ K-means clustering method to cluster radar returns into TCs. Denote TC set as \( U_1 = \{t_{c_1}^1\} \).
3. Step 2: Create a Kalman filter \( v_1^i = (x_1^i,y_1^i,vx_1^i,vy_1^i) \) for each TC \( t_{c_1}^1 = (cx_1^1, cy_1^1) \) by setting \( x_1^i = cx_1^1, y_1^i = cy_1^1, vx_1^i = 0, vy_1^i = 0 \) and \( L^i(1) = C \). The tracking model set is denoted as \( V_1 = \{v_1^i\} \).

for each frame \( k > 1 \) do

4. Step 3: Cluster radar returns into TCs, and denote the TC set as \( U_k = \{t_{c_k}^i\} \).
5. Step 4: Associate current TCs in \( U_k \) with existing tracking models in \( V_{k-1} \) using MAP based method;
   - Case 1: Tracking model \( v_{k-1}^i \) and current TC \( t_{c_k}^i \) form a matching pair (normal matching).
   - Case 2: No current TC associated with tracking model \( v_{k-1}^i \) (missing detection).
   - Case 3: Current TC \( t_{c_k}^i \) is associated with the added node \( D \) (new detection).

   Update \( L^i(k) \) based on Eq. 8. If \( L^i(k) = 0 \), delete the tracking model; Otherwise predict the target’s position based on state transition equation of \( v_{k-1}^i \).

   Create a new tracking model as step 2 does.

**4. Radar-Vision Registration**

As the CCD camera is situated directly above Delphi ESR in our system, 3D radar and camera coordinate systems are related via a rotation \( R \), swapping Y and Z of radar system, and a translation \( T \), with displacement only in Y direction. As a result, the problem of registration between 3D radar
space and 2D image plane can be easily converted to the equivalent of the camera calibration problem for generating ROIs. Note that Z coordinate values of radar returns in 3D radar space are always equal to 0, as Delphi ESR is a 2D radar sensor.

We employ the simple pinhole camera model to estimate the linear mapping $A$ between 3D camera coordinate frame and a 2D image plane. The mapping $A$ is in the form as follows.

$$A = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$  \hspace{1cm} (9)$$

In Eq. 9, $(f_x, f_y)$ and $(c_x, c_y)$ are camera focal length and optical centers, respectively. With lens model available, radar return $(R, \theta)$ will be projected to pixel $(u, v, w)$ in homogeneous coordinate by the following equation.

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = A \times [R \mid T] \times \begin{bmatrix} R \sin \theta \\ R \cos \theta \\ 0 \end{bmatrix} = M \times \begin{bmatrix} R \sin \theta \\ R \cos \theta \\ 0 \end{bmatrix}$$  \hspace{1cm} (10)$$

For our specific setup, we have $T = [0, -0.1, 0]^T$ and $R = [1, 0, 0; 0, 0, 1; 1, 0, 0]$. We employ 2D chessboard pattern to calibrate the CCD camera [6] (i.e., computing $A$), and finally obtain $M$ as follows.

$$M = \begin{bmatrix} 1611.9 & 638.5 & 0 & 0 \\ 0 & 425.9 & 1607.1 & -160.7 \end{bmatrix}$$  \hspace{1cm} (11)$$

5. Experiments

We consider two scenarios for our experiments to test the effectiveness of the developed human tracking approach. Then we present the results of our experiment to show the performance of our registration approach in locating ROIs.

5.1 Single Human Tracking

We described a human detection experiment using Delphi ESR in Section. 2. In that experiment, Delphi ESR missed the human target 26 times during the experiment. In this section, we use the human tracking approach, described in Algorithm 1, to recover from the human misses.

True human walking trajectory (red) as well as estimates from Kalman filters (blue) are both depicted in Fig. 7. We make the following observations from the result.

- Human misses of Delphi ESR are completely recovered by our tracking approach. We also calculated the mean squared error of Kalman filter estimates for this experiment. The mean squared error is equal to 0.3119m. The estimation error can be explained as follows. When the human target changes its walking direction, the Kalman filter needs some time to catch up with the target. Therefore, one conclusion we draw is that our human tracking approach has the capability to track the human target during the process.
- Our Kalman filter converges with time. That means that our measurement noise model as described in Section. 3.1 is accurate enough to describe Delphi ESR’s measurement noise.

![Fig. 7: One example of human tracking using Delphi ESR. Note that the true human walking trajectory is computed by smoothing raw radar returns in a consecutive frame window.](image)

5.2 Multiple Human Tracking

Our second experiment is conducted in dusty environment as shown in Fig. 1. During the experiment, we keep our fusion system at a fixed position. Two human targets walk in different directions, and at a specific time they meet each other and then depart away. This was done to test the effectiveness of MAP-based association approach.

Fig. 8a shows raw radar returns from human targets. We noticed that radar missed the human targets 30 times. We apply the new tracking approach, described in Algorithm 1, to recover from the human misses. Estimates of Kalman filters (i.e., cyan and blue solid lines) as well as true human walking trajectories (i.e., magenta and red dotted lines) are depicted in Fig. 8b. In this figure, black arrows indicate walking directions of two human targets. The two target meet at the bottom intersection point and then depart away from each other. From the result, we observe that the tracking approach can effectively track the targets. The mean squared estimation errors of Kalman filters for the two targets are equal to 0.4200m and 0.2591m, respectively. These are mainly caused by the sudden change in the walking directions.

Another observation is that even when two human targets are close enough, our tracking approach has the capability to differentiate and track them correctly. This can be explained by the fact that MAP-based association method ensures one-to-one mapping between current TCs and existing tracking models.

5.3 Mapping radar returns to image plane

We present an example in Fig. 9 to show the accuracy of our registration approach. This experiment is carried out in an environment where a surface miner is working in field. The human target is walking side by side at average 0.75.
m/s. In this experiment, we firstly employ range filtering to filter out radar returns beyond 20m. This is because we set a goal that we are only interested in human targets within 20m. We then utilize the approach of Algorithm 1 to track each TC. Estimated positions of each TC from Kalman filters are finally mapped to images based on Eq. 10 to generate ROIs.

We present 4 consecutive snapshots in Fig. 9. Locations of ROIs mapped from Delphi ESR returns are indicated by red circles. Registration results correspond to what we expect: radar returns from the human target and surface miner are mapped to these targets in images correctly after registration, although with small offsets. It is worth mentioning that we make the observation that the lateral position error of Delphi ESR for a typical human targets is small and consistent until the target’s lateral speed reaches close to 1.0 m/s, at which point the lateral position error gets large. This is due to beam scanning pattern of Delphi ESR. Therefore, one need to compensate the offset for ROI generation purpose when lateral velocity is larger than 1.0 m/s.

6. Conclusions

We build a radar-vision fusion system to handle human detection and tracking in dusty and smoky environments. Our system utilizes radar returns to generate ROIs and then employs a vision-based human detection technique to validate each ROI. Based on experimental observations that human misses from Delphi ESR are intermittent, and Delphi ESR’s measurement noise approximately follows Gaussian distribution, we develop a novel human tracking approach using Kalman filter to recover from human misses. In addition, we develop a radar-vision registration approach via camera calibration to map radar returns to image plane for generating ROIs. A set of real-world experiments showed that our tracking approach has the capability to accurately track human targets, and our registration approach can locate ROIs correctly when lateral velocity is smaller than 1.0 m/s.

7. Acknowledgment

The authors would like to thank Dr. Koray Celik for building the real-time data acquisition system, allowing us to use the system for our research, and providing valuable guidances on our work. Besides, the authors acknowledge Philip and Virginia Sproul Professorship funds from Iowa State University.

References


