Sensor Fusion of 3D Time-of-Flight and Thermal Infrared Camera for Presence Detection of Living Beings

Moritz Oppliger, Jonas Gutknecht, Roman Gubler, Matthias Ludwig and Teddy Loeliger
Institute of Signal Processing and Wireless Communications
ZHAW Zurich University of Applied Sciences
Winterthur, Switzerland
moritz.oppliger@zhaw.ch

Abstract—Presence detection of living beings is a key processing step in many applications. To increase the detection accuracy, an algorithm that uses camera fusion of a 3D Time-of-Flight (3D ToF) and a Long-Wave Infrared (LWIR) camera is applied and characterized here. Different sensor fusion approaches and sensor data processing schemes are compared. The algorithms are applied to a typical wildlife detection dataset and their performance is evaluated. The detection accuracy is up to 10% higher when combining 3D ToF and LWIR images compared to only 3D ToF or LWIR images. Additionally, the location and size of the living being can be determined using the 3D ToF image.

Index Terms—3D Time-of-Flight (3D ToF), Long-Wave Infrared (LWIR), Thermal Imaging, Camera Fusion, Sensor Fusion

I. INTRODUCTION

Camera traps are widely used in wildlife research, conservation and photography as well as in smart buildings, e.g. for people counting [1]–[3]. When a living being is in front of the camera trap, it triggers a high-resolution camera. The detection of the living being can be based on different technologies. Passive infrared (PIR) sensors are often used for presence detection of people and wildlife [4]–[6]. But the detection reliability by using camera traps with a PIR sensor is very sensitive to their placement [6]. Also, the sensitivity of PIR sensors is usually very high, resulting in an unwanted high false positive (FP) rate [7]. The presence detection can also be based on radar sensors [8]. The use of radar sensors in wildlife monitoring typically results in a lot of FP when there is wind or rain [9]. By using infrared light barriers to detect living beings, a detection occurs when the light barrier is crossed, limiting the camera’s triggering to a very limited space [10].

By introducing a two stage detection system, the high FP rate can be reduced. The PIR sensor triggers the second detection stage, which triggers the high-resolution camera. For the second stage, a low-resolution camera could be used. By applying a background subtraction (BS) algorithm to the images, changes as described in sec. IV can be recognized and interpreted as a living being. If the second stage detection is based on a camera in the visible spectrum, the detection works well when the illumination of the field of view (FoV) of the camera is strong enough. When the FoV is dark, the illumination would need a lot of energy and would disturb the scene due the visibility of the light.

When using a 3D-Time-of-Flight (3D-ToF) camera as a second detection stage, the field of view (FoV) has to be illuminated in the infrared spectrum at 940 nm, which is invisible to most animals. With the 3D ToF camera, there are additional information available like the location of the detection in the 3D space. Multi-path interferences can lead to significant errors in 3D ToF measurements, but advanced algorithms are available in order to resolve these limitations [11]. In addition to the 3D ToF camera, a Long-Wave Infrared (LWIR) camera is used, which, in contrast to the 3D ToF, can detect thermal radiation.

In order to achieve the fusion of the 3D ToF and LWIR images, the cameras and their relative displacement have to be calibrated. Since the LWIR and the 3D ToF camera have no overlap in the detectable spectrum, a multispectral calibration plate is used for the calibration. In stereo vision, the defined features on the calibration plate are only visible in the visible spectrum. For calibration of LWIR images a thin mask with cut out squares is proposed in [12]. The mask is placed in front of a warm object (e.g. a warm monitor), so the thermal radiation of the warm object can pass the holes and is detectable with the LWIR camera. Other multispectral calibration plates are presented in [13].

In contrast to [14], which tracks people based on fused 3D ToF and LWIR cameras, the main contributions of this paper are the new detection algorithms and the performance evaluation of these algorithms.

II. HARDWARE SETUP

A. Sensors

The technical specifications of the used cameras are summarized in Table I. The images are collected with a microcontroller and stored on a SD Card. The cameras are connected in a mechanically stable manner so that the relative displacement is constant (Fig. 1).
<table>
<thead>
<tr>
<th>Used Cameras</th>
<th>Principle</th>
<th>Resolution</th>
<th>FoV</th>
<th>Spectrum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flir Lepton 3.5</td>
<td>LWIR Image</td>
<td>160×120 px</td>
<td>50°×19°</td>
<td>8 to 14 μm</td>
</tr>
<tr>
<td>epc635</td>
<td>3D ToF</td>
<td>160×60 px</td>
<td>57°×42°</td>
<td>940 nm</td>
</tr>
</tbody>
</table>

Fig. 1. Used hardware.

Fusion performed in 3D space assigns a temperature $T$ to each point $(X_{ToF}, Y_{ToF}, Z_{ToF})$ of the 3D ToF image. A fused image with $(X_{ToF}, Y_{ToF}, Z_{ToF}, T)$ represents one sample.

B. Camera Model

In computer vision (CV) it is common to use the pinhole model for mapping points from the 3D world $(X, Y, Z)$ onto the image sensor $(x_s, y_s)$ and the mapping is calculated as following [15]:

$$\begin{pmatrix} x_s \\ y_s \\ 1 \end{pmatrix} = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \cdot \frac{1}{Z}, \quad (1)$$

where $f_x, f_y$ are the focal length of the lens in pixels and $c_x, c_y$ the center offsets, which are found through calibration (Sec. II-D). In a real camera, radial and tangential distortions occur due to the non-ideal properties of the lens and the assembly process of the sensor. The coefficients for the elimination of radial and tangential distortions are determined by calibration.

C. Transformation

A point in the FoV of the 3D ToF camera $(X_{ToF}, Y_{ToF}, Z_{ToF})$ can be transformed to the FoV of the LWIR camera $(X_{LWIR}, Y_{LWIR}, Z_{LWIR})$ with the following equation [15]:

$$\begin{pmatrix} X_{LWIR} \\ Y_{LWIR} \\ Z_{LWIR} \\ 1 \end{pmatrix} = \begin{pmatrix} R_{ToF}^{LWIR} & t_{ToF}^{LWIR} \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} X_{ToF} \\ Y_{ToF} \\ Z_{ToF} \\ 1 \end{pmatrix}. \quad (2)$$

$R_{ToF}^{LWIR}$ is the rotation matrix and $t_{ToF}^{LWIR}$ the translation vector between the 3D ToF and the LWIR camera. The matrix and the vector are determined by calibration (Sec. II-D). The LWIR image is transformed by backward transformation with a bilinear interpolation into the FoV of the 3D ToF camera.

D. Calibration

The calibration pattern has to be visible in the spectrum of the LWIR and 3D ToF camera at the same time. Therefore, a mask out of 3 mm thin Depron® with an asymmetric circle pattern is produced (Fig. 2 a). The mask is glued to an aluminium plate. To reduce unwanted reflections, the plate and the mask are coated with a matt white foil, and the holes of the mask are filled with a matt black color. Due to the black color, the pattern is visible in the amplitude image of the 3D ToF camera (Fig. 2 c). In order to make the pattern visible in the LWIR image, the aluminium plate is heated with hot water. Since Depron® has a low thermal conductivity, the circles are well visible in the LWIR image (Fig. 2 b).

Fig. 2. Images of the calibration plate with a) normal camera, b) LWIR camera, c) 3D ToF camera (amplitude image).

III. DATASET

The dataset, consisting of 218 samples, can be divided in 65 % lab samples and 35 % real-live samples, where 68 % contains a living being and 32 % are without a living being. For the lab setup, an environment with some moving grass or leaves in the wind is simulated with moving cords (Fig. 3 a). A warm toy animal is pulled through the scene to represent a dummy animal. The real world data are collected in a fallow deer enclosure (Fig. 4) and in the garden.

Fig. 3. a) Image of the used setup in the lab. b) 3D ToF distance image and LWIR image of the setup.

The dataset, consisting of 218 samples, can be divided in 65 % lab samples and 35 % real-live samples, where 68 % contains a living being and 32 % are without a living being. For the lab setup, an environment with some moving grass or leaves in the wind is simulated with moving cords (Fig. 3 a). A warm toy animal is pulled through the scene to represent a dummy animal. The real world data are collected in a fallow

Fig. 4. LWIR example image of fallow deer.
IV. DETECTION ALGORITHM

The algorithm is based on BS [16]. The background is modeled using a Gaussian distribution (mean and standard deviation) for every pixel of the 3D ToF and LWIR image. This model is estimated by an Infinite Impulse Response filter from the previous images. After performing the BS, the resulting difference image $\Delta$ is thresholded with an adaptive threshold, which depends on the standard deviation $\sigma$:

$$m(x, y) = \begin{cases} 1, & \Delta(x, y) > \sigma(x, y) \cdot \gamma \\ 0, & \text{else} \end{cases}$$

resulting in the mask $m$, where $\gamma$ is a weight to set the sensitivity. If the size of the largest connected region in the mask $m$ is larger than a threshold, a living being is detected. The fusion of the LWIR and ToF image can be realised at different stages in the processing chain as shown in Fig. 5. The two cameras are linked with an AND or an OR operation. This results in six detection algorithms for the fused images. If the AND connection is used, the thermal radiation and the distance have to change. If the algorithm uses the OR connection, only a change of one of the two features is needed for a detection.

![Fig. 5. Different approaches of the detection algorithms.](image)

### A. Image Based Fusion

The imaged based fusion analyses the LWIR and 3D ToF image separately. The region independent version evaluates the separate detection results of the LWIR and ToF images logically. The region dependent algorithm evaluates each pixel of the detection masks logically and then evaluates the size of the common mask. The possible logical operations can be an AND or an OR connection.

### B. Pixel Based Fusion

The pixel based fusion calculates a voting plane after the BS. For the OR connection, one pixel of the voting plane is calculated as following:

$$v(x, y) = \alpha_d \cdot \frac{|\Delta d(x, y)|}{\sigma_d(x, y)} + \alpha_T \cdot \frac{|\Delta T(x, y)|}{\sigma_T(x, y)},$$

For the detection, the voting plane is thresholded by a value, which results in the mask $m$. If a connected region of the mask from the thresholded voting plane is large enough, the algorithm detects a living being.

V. RESULTS AND DISCUSSION

The dataset from Sec. III was used to evaluate the algorithms (Table II). The confusion matrix and the accuracy (Acc) are used as quality criteria. The Acc is the sum of the true positives (TP) and true negatives (TN). For the comparison of the algorithms, the focus is placed on the Acc. If just the LWIR camera is used, the Acc is 6 % higher than with the 3D ToF camera alone. To detect ectothermic animals like reptiles, the LWIR camera would most probably fail. If the Algorithm fuses the 3D ToF and LWIR cameras, the logical connection type is the most important factor. By using an OR connection, the Acc is about 5 to 11 % higher compared to using a single camera, particularly if the living being is in front of a background with high variance in distance (like the lab setup). In addition, by using the 3D ToF camera, the location and size of the living being can be determined.

![Table II: Results of the algorithms in percent.](image)

VI. CONCLUSION

We propose different new algorithms for living being detection by using 3D ToF and LWIR images. The detection Acc by using fused 3D ToF and LWIR images can be increased by about 5 to 10 % compared to using a single camera, particularly if the living being is in front of a background with high variance in distance (like the lab setup). In addition, by using the 3D ToF camera, the location and size of the living being can be determined.

ACKNOWLEDGMENT

This work has been funded by the Swiss National Science Foundation SNSF (PT00P2_198986).
REFERENCES


