Audio-Visual System for Cooperative Localization of Small Aerial Robots

José Eduardo Fernandes Rosa
jose.rosa@ist.utl.pt

Instituto Superior Técnico, Lisboa, Portugal
May 2019

Abstract

Employing small size aerial robots, acting as mobile airborne sensors, to work alongside ground robots can be extremely useful in many different robotic missions. Due to the strict constraints in terms of size, weight, 3D coverage, processing power, and power consumption, there are not many technological possibilities for performing independent self-localization for such tiny robots. This thesis describes an audio-visual system for cooperative localization to robustly estimate the position of a small aerial robot with respect to a ground robot. Experimental results with a 40g quadrotor assess the performance of the system and demonstrate the reliability gained through the fusion of acoustic with visual information with an extended Kalman filter. The audio-visual system resorts to a microphone array and a Microsoft Kinect v2 camera for sound and visual measurements, respectively, achieving an accuracy mean absolute error of \(0.014m\) for 3D position estimation.

Keywords: Audio-Visual Localization, Micro Aerial Robots, Sensor Fusion, Microphone Array, Ground-Aerial Cooperation

1. Introduction

Small aerial robots have seen their use increased in recent years due to its very lightweight and small size properties. These characteristics provide the robot very high mobility and agility which can be of great utility in various robotic missions. Nonetheless, these very same properties also pose limitation, since the small size and low weight impose strict constraints to the on-board equipment carrying capabilities of such tiny robots. For this reason, there are not many technological solutions available that could satisfy these constraints and allow the self-localization of small aerial robots. Contrarily, ground robots are geared with high-end technology sensors without major restrictions on weight or size. Therefore, this thesis proposes an audio-visual solution to be applied through cooperation where the ground robot sensors, which provide accurate measurement tools, can be utilized for detection and relative localization of small aerial robots. This solution would be of great interest for both parts as it would extend the capabilities of the ground robot while performing the difficult task of localization of tiny aerial robots. The system fuses visual and acoustic sensor data with other sensory information in real time, by employing a Microsoft Kinect v2 depth camera, Figure 1(a) and an eight microphone array, Figure 1(b), allowing robust and reliable pose estimation of the micro aerial robot Crazyflie 2.0, Figure 1(c).

The system can be useful for many robotic solutions and in particular for teams of ground and aerial robots. Further and more concrete applications for the proposed system are in damaged buildings [1] and inaccessible areas for the ground robot [2, 3].

2. Related Work

There exist very few technological possibilities that could provide onboard localization while satisfying the strict constraints of such small robots in terms of weight, size, power consumption, processing power, three-dimensional coverage [4–6]. In [5] is used a 4.7g camera to perform onboard vision based localization of a 40g quadrotor. However, the solution was highly dependent on illumination and available visual features in the environment.

An alternative to on-board systems is to use an external positioning system such as motion tracking cameras [7, 8] or wireless beacons [9, 10]. However, such systems are required to be deployed in advance, which is not always efficient or even practical.

While vision based sensors are shown to be very successful and widely used in many robotic solutions for localization and tracking targets [11–13], they suffer some limitations such as the dependency on illumination, visual contrast, and the limited field of view. For this purpose, acoustic sensors are used in this work to complement the
vision-based information and enhance the robustness of position estimations. This is achieved by exploiting the already available sound waves from the engines of the micro aerial robots to extract additional pose information that does not suffer the same type of limitations as with the vision sensors. Sound-based localization of micro aerial robots, using microphone arrays, is usually focused on the direction of arrival (DOA) rather than the 3D position [14, 15]. These systems do not grant the same level of localization accuracy but have different limitations.

As to multi-sensor fusion, Extended Kalman filter (EKF) is the most common sensor fusion technique in robotics. EKF is a reliable approach for real-time state estimation [16] that is widely utilized in position estimation of aerial robots [10, 17, 18].

3. Vision based position estimation

Cameras are very versatile and frequently present in robots as they are utilized in numerous different tasks. RGB-D or depth cameras, in particular, have the ability to estimate three-dimensional position by employing a depth sensor alongside the regular camera.

3.1. Image Acquisition

While working in the Robotic Operative System (ROS) conversion of the sensor data obtained from the RGB-D camera to appropriate ROS messages is necessary. For the chosen camera, Microsoft Kinect v2, this was achieved through the existing libraries iai_kinect2 [19] and libfreenect2 [20]. For this thesis, rectified images with quarter FullHD resolution (960x540) were used.

3.2. Aerial Robot Detection

For the detection phase, a simple strategy was used where the micro aerial robot was rigged with a red ping pong ball on top of the frame, Figure 3.2(a), which allowed shape and colour detection, Figure 2(b). Afterwards, pixel coordinates of the ball centre were computed.

3.3. Visual Position Estimation

After detecting the ball centre in two dimension pixel coordinates \((x_p, y_p)\), the pixel depth \((d_p)\), in metres, is determined. Using these values and the camera calibration matrix, it is possible to compute the real values \((x_r, y_r)\), in metres, of the position with respect to the camera.

As a convention, depth cameras axes are according to Figure 3(a), however, in this work, the position obtained by the camera, was converted to a more intuitive axes system, in correspondence with the world coordinate frame, Figure 3(b).

3.4. Experimental Setup

To evaluate the camera position estimation measurements it was crucial to design suitable experiments as well as to have a site that allowed the tests to be conducted. The test environment chosen was the Mbot test bed located in the ISR laboratory, Figure 4. The laboratory is equipped with 12 Optitrack motion capture system cameras, that are utilized as ground truth.

As a way to compute and evaluate the position estimation of the Crazyflie, four reflective markers were added to its body frame, Figure 5.

3.1. Image Acquisition

While working in the Robotic Operative System (ROS) conversion of the sensor data obtained from the RGB-D camera to appropriate ROS messages is necessary. For the chosen camera, Microsoft Kinect v2, this was achieved through the existing libraries iai_kinect2 [19] and libfreenect2 [20]. For this thesis, rectified images with quarter FullHD resolution (960x540) were used.

3.2. Aerial Robot Detection

For the detection phase, a simple strategy was used where the micro aerial robot was rigged with a red ping pong ball on top of the frame, Figure 3.2(a), which allowed shape and colour detection, Figure 2(b). Afterwards, pixel coordinates of the ball centre were computed.

3.3. Visual Position Estimation

After detecting the ball centre in two dimension pixel coordinates \((x_p, y_p)\), the pixel depth \((d_p)\), in metres, is determined. Using these values and the camera calibration matrix, it is possible to compute the real values \((x_r, y_r)\), in metres, of the position with respect to the camera.

As a convention, depth cameras axes are according to Figure 3(a), however, in this work, the position obtained by the camera, was converted to a more intuitive axes system, in correspondence with the world coordinate frame, Figure 3(b).

3.4. Experimental Setup

To evaluate the camera position estimation measurements it was crucial to design suitable experiments as well as to have a site that allowed the tests to be conducted. The test environment chosen was the Mbot test bed located in the ISR laboratory, Figure 4. The laboratory is equipped with 12 Optitrack motion capture system cameras, that are utilized as ground truth.

As a way to compute and evaluate the position estimation of the Crazyflie, four reflective markers were added to its body frame, Figure 5.

(a) Reflective marker
(b) Crazyflie with four reflective markers

Figure 5: Reflective markers utilized in the Optitrack motion capture system.
3.4.1 Flight Tests

The position estimation was evaluated during a series of tests, made in-flight, which were designed to move the robot in every direction. These tests were conducted with the assistance of the Crazyflie ROS package \cite{21} which allowed a route of pre-determined waypoints to be set.

To evaluate the camera performance, two different sets of points are presented. The first maintains the aerial robot within the camera field of view, Figure 6(a) whilst the second pushes the robot outside the camera FOV limits as well as outside the depth sensor optimal range, Figure 6(b). The first waypoint route takes approximately 25 seconds while the second runs over 35 seconds.

![Trajectory](image)

Figure 6: Sequence of aerial robot detections projected on RGB image obtained from the Kinect camera.

3.5. Results

Figures 7 and 8 present the measurements from the Microsoft Kinect v2 camera, with respect to the world origin, in three-dimensional coordinates during the 2 experimental routes.

From Figures 7 and 8 is observed that both experiments present very accurate results. Nevertheless, in the second set, an interval of inaccurate measurements (around 27s) is present. This happens when Z is close to 0 and Y is close to \(-1\) which, in practice, means the aerial robot was near the ground and in close proximity to the camera - the lowest point in Figure 6(b). The fact the robot is so close to the camera, damages the depth measurements in Figure 8(b).

Since the measurements presented are obtained in real-time there is a time delay associated with them. This delay is observable since both curves are slightly shifted, and is estimated to be close to 70ms. The cause for the delay is the update rate of the camera (30Hz) and the computation cost of the position estimation algorithm.

Alongside the position plots, error histograms for each axis are presented - Figure 9. From these plots, is observed a normal distribution of the error, with Gaussian parameters in Table 1.

In conclusion, Microsoft Kinect v2 camera is a reliable and accurate source of measurements and can safely be utilized for robot localization. The errors are small, order of millimetres, and present a Gaussian distribution which will be of extreme importance in the use of the Kalman Filter. The mean absolute error (MAE) for 3D position estimation in this system is 0.013m.
3.6. Limitations and Improvements

From the tests conducted some limitations and potential improvements of the work are outlined:

- The detection of the Crazyflie could be improved. There are methods available, such as feature detection or background processing, which could be used to remove the red ball and make the robot lighter. It would also result in a more robust detection for the real case where the environment colours are not restricted.

- With respect to the field of view of the camera, it could be implemented some conditions in the code which condition the movement of the aerial vehicle to avoid reaching out of the camera FOV.

4. Audio Based Direction of Arrival Estimation

In this thesis, audio based localization is chosen by performing independently of vision based positioning methods. The design of an audio based DOA algorithm, to localize the aerial robot by its natural flying sound, using an eight microphone array, is described.

Direction of Arrival is divided into two angular components: the azimuth $\alpha$ (Equation 1) and the elevation $\beta$ (Equation 2). The mentioned equations derive DOA angles from the 3D coordinates of the audio source. Azimuth is measured in the $xy$ plane and ranges from $-\pi$ to $\pi$ while elevation is the angle between the $xy$ plane and $z$ axis and ranges from $-\pi/2$ to $\pi/2$.

$$\alpha = \arctan(x/y)$$ (1)
$$\beta = \arctan(z/\sqrt{x^2 + y^2})$$ (2)

4.1. Microphone Array

Microphones are commonly present in ground robots as they are necessary for speech recognition. However, most microphone solutions are not suitable for DOA estimation - either by not having the necessary number of microphones or by not having the microphones positioned in a way to allow both azimuth and elevation to be unambiguously estimated. This section describes the microphone solution developed, enabling azimuth and elevation to be computed.

It was decided to build a microphone array composed of eight microphones, which could be mounted on a ground robot. With eight sensors, it is possible to design a geometry arrangement which, theoretically, allows unambiguous measurements both in the horizontal and in the vertical planes (azimuth and elevation). In addition, the array was also expected to be homogeneous. Therefore, it was decided that the microphone array should be shaped like a cube.

4.2. Localization Method

The Beamforming technique Steered-Response Power Phase Transform (SRP-PHAT) was used for sound localization as it provides the best results. It was applied with pyroomacoustics python package [22] which is a library developed for sound testing purposes. It allows easy comparison between different localization methods while supporting custom resolutions and angle spectrum.

SRP-PHAT method is divided into two steps [14]. Firstly, it computes the coherence measuring unit $C_{ij}$ (Equation 3) which evaluates the correlation between each pair of microphones.

$$C_{ij}(\tau) = \frac{\text{FFT}^{-1}\left(\text{FFT}(p_i(n)) \cdot \text{FFT}^*(p_j(n))\right)}{W}$$ (3)

where $\text{FFT}$ corresponds to the Fast Fourier Transform, $\text{FFT}^{-1}$ to its inverse, $\text{FFT}^*$ to its complex conjugate, $p_i(n)$ is the discrete signal sequence from microphone $i$ and $\tau$ is the correlation lag. $W$ corresponds to the spectral weighting function PHAT (Equation 4) which improves localization performance by giving equal importance to all frequencies and achieving greater correlation peaks.

$$W = |\text{FFT}(p_i(n))| |\text{FFT}(p_j(n))|$$ (4)
Secondly, after computing the coherence unit for each pair of microphones, the search for the acoustic source direction \( \vec{b}_m \) is made. This computation (Equation 5) aims at maximizing the sum of \( C_{ij}(\tau) \) where the time delay \( \tau \) corresponds to the direction of the search. This time delay is obtained from the microphone’s spatial positions.

\[
\vec{b}_m = \arg\max \sum_{i,j} C_{ij}(\tau_{bij}) \tag{5}
\]

4.3. Experimental Setup

In the experimental context and real application of the acoustic system, a full 3D directional scan is not always required. Therefore, the SRP-PHAT algorithm’s search spectrum could be focused around the region where the acoustic target operates, in order to improve its computational speed and detection resolution, resulting in better real-time measurements. Moreover, if required, the search spectrum of SRP-PHAT for azimuth and elevation could be updated in real-time, based on the last measurements obtained.

`pyroomacoustics` python library, provides all the necessary tools to easily compute real-time DOA estimation. It requires only knowledge of the microphone distribution (position in space) and the data feed from each one. Furthermore, it also provides customization variables to define the angles’ search spectrum, the resolution of the measurements and the frequency band of search. The band was defined by recording the aerial robot’s produced sound and ambience noise, separately, and posteriorly compute its frequency difference.

4.3.1 Measurement Acquisition

Identically to the position estimation through depth camera, the Optitrack motion capture system was utilized as ground truth. It should be noted that the measurement readings from the DOA estimation are in azimuth and elevation while the readings from the motion capture system are 3D coordinates. This fact requires the Optitrack measurements to be converted to the microphone array frame since azimuth and elevation cannot be converted to 3D coordinates.

4.4. Results

Figures 11 and 12 show the obtained values of azimuth and elevation angles.

In these figures, is seen that the measurements follow the gist of the Crazyflie movement yet there is a considerable number of outliers as well as scattered measurements. Additionally, in comparison with vision based plots, the number of measurements is considerably lower. This is a direct consequence of the publish rate which is around 15Hz which, despite not being as high as desired, allows real-time estimation.

Further analysis of the measurements is made with the aid of error histograms presented in Figure 13 and compounding results in Table 2. The values follow a Gaussian distribution. By comparing the values from Table 2 a slight bias in the measurements, which is considerably higher for elevation, is observed. This bias is estimated to be present due to ground reflections and environment noise. The standard deviation values for azimuth and elevation are similar which demonstrates the microphone array is homogeneous, as intended. Despite this fact, these values have considerable deviations which are higher than desirable. Considering the worst test, the MAE for azimuth and elevation is 7.55 deg and 8.20 deg, respectively.

\[
\begin{array}{l|ll|ll}
 & \text{Set 1} & \text{Set 2} \\
\hline
\alpha & 0.31477 & 5.69256 & -0.30104 & 7.55245 \\
\beta & 3.17450 & 6.29400 & 1.39864 & 8.07447 \\
\end{array}
\]

Table 2: Mean and standard deviations of the Gaussian fit for each audio histogram error measurements - values in degree.
4.5. Limitations and Improvements

The audio based solution for localization of small aerial robots poses some limitations and possible improvements in the future:

The current solution does not have any kind of sound separation technique. This means the solution is not robust when environment noise, in the frequency band of the small aerial robot, is present. Sound separation methods could be applied as well as a more thorough detection process of the robot produced sound.

As detected in the plots, there are considerable deviations which could be due to the effect of sound reverberation. This issue can be further tested and possibly improved by assessing the environment with greater detail and if required implement reverberation improvement algorithms.

The publishing frequency of SRP-PHAT, proved to be the fastest, with 15Hz rate. While 15Hz is enough for real-time applications, higher values would bring greater estimations. The use of a powerful Graphics Processing Unit (GPU) would likely improve this parameter.

5. Extended Kalman Filter

It was designed an extended Kalman filter, for sensor fusion of visual and acoustic data. EKF allows the use of non-linear equations to describe both the system model and the measurement model.

5.1. Sensor Data

For more complete tasks, in aerial robot localization, is usually necessary to also have information about the orientation (roll $\varphi$, pitch $\theta$, and yaw $\psi$). Considering visual and audio localizations do not provide orientation measurements, it was decided to include the Inertial Measurement Unit sensor to further improve EKF’s prediction.

Figure 14 and Table 5.1 show the IMU orientation values. Figure 14(d) shows a mean systematic error, due to initialization procedures, as the robot resets the IMU every time it turns on.

<table>
<thead>
<tr>
<th>Orientation Angle</th>
<th>$\mu$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varphi$</td>
<td>1.79639</td>
<td>0.88209</td>
</tr>
<tr>
<td>$\theta$</td>
<td>-1.74359</td>
<td>0.96509</td>
</tr>
<tr>
<td>$\psi$</td>
<td>-1.35910</td>
<td>0.66315</td>
</tr>
</tbody>
</table>

Table 3: Gaussian fit parameters for the error distribution of the IMU - values in degree.

5.2. Equations

5.2.1 Problem Description Equations

This section is dedicated to the theoretical component of the extended Kalman filter and its equations. The presented representations follow the well-known book Probabilistic Robotics [23].

The general equations to describe the extended Kalman filter are Equation 6 and 7 which define the state transition model, $u_t$ is the control input and $E_t$ is the noise, assumed to be Gaussian.

$$X_t = g(u_t, X_{t-1}) + E_t$$  \hspace{1cm} (6)

following the same structure as the previous equation, where $Z_t$ is the measurement vector at time instance $t$, $h$ is a non-linear function describing the measurement model and $\delta_t$ is the measurement noise which is considered Gaussian as well.

For the purpose of describing and predicting the small aerial robot’s position, in this thesis, the extended Kalman filter state vector $X$ is composed by the position $(x, y, z)$, the orientation $(\varphi, \theta, \psi)$, the linear velocities $(\dot{x}, \dot{y}, \dot{z})$ and the angular velocities $(\dot{\varphi}, \dot{\theta}, \dot{\psi})$.

$$X_t = [x, y, z, \varphi, \theta, \psi, \dot{x}, \dot{y}, \dot{z}, \dot{\varphi}, \dot{\theta}, \dot{\psi}]$$  \hspace{1cm} (8)

Since the control vector is not present in this work ($u_t = 0$), the non-linear function depends only on the previous state, $g(u_t = 0, X_t)$. For the system model a simple kinematics model with a rotation matrix was used, to describe the small aerial robot behaviour in the world frame.

The measurement vector is composed of the sensor data obtained. In the context of this thesis the sensors and their corresponding measurements are:

Microsoft Kinect v2, which provides the three-dimensional position of the Crazyflie, relative to the global frame. The $x$, $y$ and $z$ coordinates are the parameters given to the measurement vector.

Crazyflie’s Inertial Measurement Unit provides the orientation angles $(\varphi, \theta, \psi)$ and angular velocities $(\dot{\varphi}, \dot{\theta}, \dot{\psi})$ to the measurement vector, relative to the robot’s frame.

![Figure 14](https://via.placeholder.com/150)
Microphone Array, which provides DOA measurements (α and β from the measurement vector) of the Crazyflie with respect to the microphone array centre point.

The measurement model is then expressed by:

\[ Z_t = [x, y, z, \phi, \psi, \theta, \dot{\phi}, \dot{\psi}, \dot{\theta}, \alpha, \beta] \]  

(9)

From Equation 7 is seen that the majority of the variables, with the exception of azimuth and elevation, is present in the state vector. Therefore, the non-linear function \( h \) is direct except for the last two terms which follow Equation 1 and 2.

5.2.2 Extended Kalman Filter Equations

Adopting the terminology from the literature, the EKF belief \( \text{bel}(X_t) \) is expressed by a mean \( \mu \) and a covariance \( \Sigma \).

In this work, the extended Kalman filter algorithm is divided into three steps:

**Prediction step** computes the system expected mean \( \mu_t \) and expected covariance \( \Sigma_t \) based on the previous estimation and system model behaviour.

\[ \mu_t = g(\mu_{t-1}) \]  

(10)

\[ \Sigma_t = G_t \Sigma_{t-1} G_t^T + R_t \]  

(11)

where \( G \) is the Jacobian of \( g \), and \( R \) is the process noise covariance.

**Matching Step**, where the measurements are compared to the predicted value in order to check if they are within a certain threshold. This step ensures EKF convergence as well as discards any outliers. This step is applied separately to the visual and the acoustic measurements to allow independent fusion of measurements.

\[ v_t = Z_t - \hat{Z}_t \]  

(12)

\[ S = H_t \Sigma_t H_t^T + Q_t \]  

(13)

\[ v_t^T S^{-1} v_t < \gamma \]  

(14)

where \( Z \) is the obtained measurements, \( \hat{Z} \) are the predicted measurements (based on the state), \( H \) the Jacobian matrix of the measurement model function \( h \), \( Q \) the measurements error covariance matrix and \( \gamma \) the rejection threshold.

**Update step** aims to correct the prediction made using the measurements observed if those were accepted by the matching step.

\[ K_t = \Sigma_t H_t^T (H_t \Sigma_t H_t^T + Q_t)^{-1} \]  

(15)

\[ \mu_t = \hat{\mu}_t + K_t (Z_t - h(\hat{\mu}_t)) \]  

(16)

\[ \Sigma_t = (I - K_t H_t) \Sigma_t (I - K_t H_t)^T + K_t Q_t K_t^T \]  

(17)

where \( K \) is the Kalman gain. Normally the update step Equation 17 would be \( \Sigma_t = (I - K_t H_t) \Sigma_t \) but it was decided to add the Joseph form \( [(I - K_t H_t)^T + K_t Q_t K_t^T] \) in order to maintain filter stability when the Kalman gain is not the optimal gain as explained in [24].

6. Results

Three distinct experiments, for different real cases, are presented and their results analysed. The plots presented show the values obtained by the designed filter with and without the audio data to demonstrate the audio effect on the filter. Visual measurements were plotted as well since knowing when visual data is not accurate is of great importance.

6.1. Inaccurate Visual Data

This test presents a predefined route similar to the previously utilized. This test serves mainly to present the results of the matching step applied in the filter. This step allows bad measurements to be rejected as seen in Figure 15 around the 26-second mark. Inaccurate visual measurements are discarded by the filter and an approximation of real values is achieved due to the filter system equations.

Both filters have outstanding performance when visual measurements are accurate. Nonetheless, the EKF with audio data outperforms the no audio data filter.

Figure 17 presents the DOA measurements accepted by the matching step during the test. From these plots is seen that there are substantially fewer measurements when compared to the plots shown in Section 4.4 but they are also closer to the real values.

6.2. False Visual Measurements

In contrast to previously described experiments, this one is made while teleoperating the aerial robot with no es-
Table 4: Mean and standard deviations of the Gaussian fit for each EKF histogram error measurements - values in meter.

<table>
<thead>
<tr>
<th></th>
<th>EKF</th>
<th>No audio EKF</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>-0.00474</td>
<td>0.01999</td>
</tr>
<tr>
<td>Y</td>
<td>-0.01514</td>
<td>0.06006</td>
</tr>
<tr>
<td>Z</td>
<td>0.01199</td>
<td>0.03991</td>
</tr>
</tbody>
</table>

Established route. In addition, it was placed a static red object, red circle in Figure 18, in order to produce false measurements by the visual sensor. These measurements will be easily detected in the plots as they correspond to the same point over time, around (2.1; 2.1; 1.1).

Figure 16: EKF error histograms for audio and non-audio data solutions of the waypoint test.

Figure 17: Direction of Arrival measurements accepted by the matching step of the EKF during the waypoint test.

Figure 18: Small aerial robot projection in the camera image during the first teleoperation test.

Figure 19 presents the position estimation obtained by the EKF. Both audio and non-audio filters are presented and is possible to deduct that both solutions are capable of detecting the false measurements and reject them, but the non-audio is considerably less. Another important aspect is the recovery time, the projected EKF manages to recover from false measurements much faster than the solution without sound, as seen in the 10 to 25 seconds interval.

Figure 20 and Table 5 present the error histograms of both audio and non-audio EKF solutions. It is understandable that while both have small mean errors ($\mu$) the standard deviation ($\sigma$) of the solution without audio data is much higher, reaching almost 1m for the X axis. After analysing both position estimation and histogram plots it is possible to conclude that the non-audio solution would not be applicable, in reality, for long time intervals of inaccurate visual measurements.

Figure 20: EKF error histograms for audio and non-audio data solutions of the first teleoperation test.

6.3. Visual Data Gaps

This experimental test was made by teleoperation of the small aerial robot, similarly to the previous one. In this test, the aerial robot exited the camera FOV for a considerable amount of time. From its position in the camera
Table 5: Mean and standard deviations of the Gaussian fit for each EKF error histogram during the first teleoperation test - values in metre.

<table>
<thead>
<tr>
<th></th>
<th>EKF</th>
<th>No audio EKF</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>0.07495</td>
<td>-0.09478</td>
</tr>
<tr>
<td></td>
<td>0.29875</td>
<td>0.92257</td>
</tr>
<tr>
<td>Y</td>
<td>-0.09793</td>
<td>0.06818</td>
</tr>
<tr>
<td></td>
<td>0.36075</td>
<td>0.66229</td>
</tr>
<tr>
<td>Z</td>
<td>0.00851</td>
<td>-0.07589</td>
</tr>
<tr>
<td></td>
<td>0.25449</td>
<td>0.45626</td>
</tr>
</tbody>
</table>

image, Figure 21, is seen that the robot leaves the camera view from the top. This test serves to evaluate the EKF behaviour to extended periods of visual measurement gaps.

Figure 21: Small aerial robot projection in the camera image during the second teleoperation test.

Starting with the position estimation, Figure 22, a major gap in visual data acquisition is observed during the approximate time interval of 7 to 13 seconds and two smaller gaps, but still relevant, in the time interval of 16 to 18 and 27 to 29 seconds. From the plots, is clear that the more complete Kalman filter outperforms the non-audio one. During the refereed gaps, the audio-visual EKF tends to follow the real position of the robot while the visual only variant drifts from the real position being only able to recover when the visual data is back online.

Error histograms, for the second teleoperation experiment, were plotted in Figure 23. In conjunction with Table 6 is possible to assess the position estimation in a situation of major loss of visual data.

Table 6: Mean and standard deviations of the Gaussian fit for each EKF error histogram during the second teleoperation test - values in metre.

<table>
<thead>
<tr>
<th></th>
<th>EKF</th>
<th>No audio EKF</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>-0.00010</td>
<td>-0.11015</td>
</tr>
<tr>
<td></td>
<td>0.10231</td>
<td>0.34160</td>
</tr>
<tr>
<td>Y</td>
<td>0.03777</td>
<td>0.02040</td>
</tr>
<tr>
<td></td>
<td>0.14076</td>
<td>0.08637</td>
</tr>
<tr>
<td>Z</td>
<td>0.00803</td>
<td>-0.01322</td>
</tr>
<tr>
<td></td>
<td>0.08202</td>
<td>0.15068</td>
</tr>
</tbody>
</table>

Table 6: Mean and standard deviations of the Gaussian fit for each EKF error histogram during the second teleoperation test - values in metre.

6.4. Final Considerations
From the previous figures, it was possible to evaluate the proposed EKF in real application situations. It was, however, not possible to evaluate its real performance when under favourable conditions. The previous histograms had its Gaussian fit present and was seen that the data did not accurately follow a Gaussian distribution. This is due to the fact that in the whole extension of the tests, the data was not always accurate. For this reason, a final histogram, Figure 24 and Table 7, is presented, for the case of favourable application conditions, where the visual data is always accurate. In these figures is seen that the complete EKF still presents better results than the no audio version and follows a Gaussian distribution. The proposed EKF had a mean absolute error of 0.014m.

Figure 22: Position estimation of the small aerial robot by the EKF developed during the second teleoperation test.

Figure 23: EKF error histograms for audio and non-audio data solutions of the second teleoperation test.
Figure 24: EKF error histograms under favourable application conditions.

Table 7: Mean and standard deviations of the Gaussian fit for each EKF error histogram under favourable application conditions - values in metre.

7. Conclusions

This thesis proposes an audio-visual system for cooperative localization of micro aerial robots. The system relies on an RGB-D camera, for visual position estimation, and a microphone array, for direction of arrival estimation. The use of two distinct systems increases the robustness of the localization algorithm.

Real application problems were simulated, where the proposed solution presented satisfactory results, which suggest its viability. An MAE of 0.014 m was achieved during favourable circumstances. This MAE value is very promising which leads to the conclusion that very accurate localization can be performed using the developed system.

7.1. Achievements

This thesis achievements are:

Development of an RGB-D camera visual localization system for integration in a sensor fusion solution with a MAE of 0.013 m for 3D position estimation.

Development of an audio based direction of arrival estimation system, composed of cube-shaped designed microphone array, for integration in a sensor fusion solution. Finally, the acoustic performance was assessed where the system obtained satisfactory, yet suboptimal results. Its MAE is 7.55 deg and 8.20 deg for azimuth and elevation, respectively.

Development of an audio-visual sensor fusion algorithm using an extended Kalman filter. The filter fuses data from the two previously described systems, as well as data from the micro aerial robot IMU. The projected system is to be employed in ground-aerial cooperation. By employing sensor fusion, it was achieved a more robust position estimation solution. The obtained MAE is 0.014 m which is comparable with the most accurate state of the art techniques, such as the use of motion capture systems which obtain < 0.02 m accuracy [8].

7.2. Future Work

The work made in this thesis could be further improved and new applications could emerge. From the improvement perspective, already mentioned issues should be tackled, such as the detection algorithms in both vision and audio systems.

For visual detection, a more robust method could be applied and instead of the red ball, visual features or background processing could be employed. Feature detection is based on visual features present in the micro aerial robot, which can be detected for position estimation. Background processing, is another alternative, which computes differences in the sequential images, allowing to detect the moving aerial robot.

Regarding audio detection, a new module could be introduced to filter the aerial robot sound. This could be achieved by comparing a template sample of the produced sound with the received. Further study on this system could also potentiate improvements in reverberation.

A final improvement to this work would be the use of a more powerful CPU + GPU in order to achieve faster publishing rates and further reducing the delay.

In respect to future applications, the first would be to implement the audio-visual system in a ground robot and assess its performance in domestic applications.

Possible application emerging from this system could be the localization of multiple aerial robots simultaneously or even localizing non-cooperative noise emitting platforms (i.e. IMU or similar sensor data is not available).

References


