An IR Local Positioning System for Smart Items and Devices

Erwin Aitenbichler, Max Mühlhäuser
Telecooperation Group, Department of Computer Science, Darmstadt University of Technology
erwin@informatik.tu-darmstadt.de

Abstract

This paper describes IRIS-LPS (InfraRed Indoor Scout), an optical infrared local positioning system. The tracked objects carry active tags that emit infrared signals which are received by a stationary mounted stereo-camera. The system is based on cheap off-the-shelf components, is easy to deploy, and features a large range of coverage. It is capable of tracking a large number of tags without significant performance impact, since the sampling rate remains constant with an increasing number of tags. A test installation of the system has been evaluated in a lecture hall. The positioning system is utilized by our audio-centric user terminal called Talking Assistant. This and other application scenarios are described at the end.

1. Introduction

The global positioning system (GPS) and wireless enhanced 911 (E911) services [3] are systems in place to determine the location of mobile users. However, these systems cannot provide accurate indoor location information and GPS is strictly limited to outdoor use. Recently, there is also growing interest in accurate position finding technologies for indoor use.

In the area of ubiquitous computing, devices are increasingly designed to be "smart" by making them aware of their context. Location is one of the most interesting physical context properties. Many other physical or situative properties depend on location and/or time. This makes indoor positioning systems an important foundation for prototyping items and devices in ubiquitous computing research.

Commercial applications include asset tracking and resource management [11]. Residental and nursing homes want to track people with special needs outside visual supervision [8]. In public safety and military applications, location systems are used for tracking prisoners.

1.1. Related work

A survey of location systems for mobile computing is given in [7]. In the following discussion, the scope is limited to tag-based systems that provide absolute location information. The current location systems can then be classified by the underlying transmission medium as follows:

RF: RF-based systems currently offer an accuracy of about 1-3 meters [7]. They do not require a direct line of sight. However, their accuracy is significantly degraded by multiple path and fading effects. Our experience shows that the measurement result can even be influenced by varying the number of people standing close to a tag. RF-based systems are suitable for certain industrial applications. In office-like environments, they often cannot distinguish between different rooms. The state-of-the-art and difficulties of RF-based technology are described in [10].

Ultrasound: In the Active Bat [7] system, tags emit an ultrasonic pulse to a grid of ceiling-mounted receivers. The system can locate tags to within 9cm of their true position. Like RF-based approaches, ultrasound-based systems suffer in their accuracy from reflections and obstacles between senders and receivers. Deploying the Active Bat system requires laying out a grid of sensors on the ceiling. This task is rather complex and requires a precise placement of sensors.

Magnetic: Magnetic tracking is commonly used in VR and motion capture applications. It offers a high resolution but is limited to a small and precisely controlled environment.

Infrared: A major disadvantage of IR optical systems is that a direct line of sight between sender and receiver is required. However, if the emitters can be mounted at exposed places, like on the Talking Assistant Headset [1], this restriction does not severely limit their use.

Commercially available IR optical tracking systems such as Optotrak [9] and Firefly [4] mainly target medical and motion capture applications. Compared to these systems, our solution supports a considerably larger field of view and higher range at the cost of reduced accuracy. Furthermore, our system is based on cheap off-the-shelf available components.

Our system consists of a number of infrared emitting tags and a stationary mounted stereo camera. This stereo

camera measures the angle of arrival of the emitted light rays at two different points in space. The spatial positions of the tracked tags are then calculated by triangulation. Because the properties of optical cameras are far from optimal, this theoretically simple principle turns out to be quite complex in practice.

2. Camera model and calibration

The camera model describes the geometric and optical characteristics (intrinsic parameters) and the 3D position and orientation of the camera frame relative to a certain world coordinate system (extrinsic parameters). The process of determining these parameters is called camera calibration.

Several different models and calibration techniques are presented in literature. The model we use is strongly based on [2] and [6]. It is briefly described in the following sections to point out the differences and which parameters we chose.

Camera intrinsic parameters

This model for the non-linear camera distortions was first introduced by Brown in 1966 and called "Plumb Bob" model (radial polynomial + "thin prism").

Let P be a point in space with the coordinates $[x_c y_c z_c]$ in the camera reference frame. By using the pinhole model, this point is projected onto the image plane. The resulting normalized image coordinates $[x_n, y_n]$ are given by

$$x_n = \frac{x_c}{z_c},\ y_n = \frac{y_c}{z_c}$$

The pinhole model expresses the relationship between object and image coordinates in an "ideal" camera. However, it is too inaccurate for most computer vision applications, especially for positioning systems. It is necessary to perform systematic corrections on the distorted image coordinates. The most commonly used correction is for the radial lens distortion which causes an image point to be displaced depending on its distance r to the image center. After including the radial lens distortion with the coefficients $kc=[kc_1,kc_2,kc_3]$, the new normalized point coordinates $P_r[x_r y_r]$ are defined as follows:

$$x_r = x_n (1 + kc_1\ r^2 + kc_2 r^4 + kc_3 r^6)$$
$$y_r = y_n (1 + kc_1\ r^2 + kc_2 r^4 + kc_3 r^6)$$

Another error component is thin prism distortion. It arises from imperfect lens manufacturing and camera assembly. This type of distortion can be adequately modeled by the adjunction of a thin prism to the optical system, causing additional amounts of radial and tangential distortions [6]. The tangential distortion with the coefficients $pc=[pc_1,pc_2]$ is modeled as:

$$x_t = x_r + 2\ pc_1\ x\ y + pc_2 (r^2 + 2 x^2)$$
$$y_t = y_r + pc_1\ (r^2 + 2 y^2) + 2\ pc_2\ x\ y$$

The corrected normalized image coordinates can now be transformed into pixel coordinates by scaling and translation. The final pixel coordinates $[x_p y_p]$ of the projection of point P on the image plane are

$$x_p = fc_x\ x_t + cc_x$$
$$y_p = fc_y\ y_t + cc_y$$

fc_x and fc_y are the focal length expressed in units of horizontal and vertical pixels. The ratio fc_x / fc_y is called aspect ratio. It is equal to 1 if the pixels in the camera's CCD sensor are square. The skew coefficient α_c has been left out in our model, we assume rectangular pixels.

cc_x and cc_y express the coordinates of the optical center axis in units of pixels, which is also called principal point.

Camera extrinsic parameters

The extrinsic transformation describes the mapping from an arbitrary world coordinate system to coordinates in the camera reference frame. It is defined by the rotation vector ωc and the translation vector Tc. Tc is the position of the camera's projection center in world coordinates. The projection center is also the origin of the coordinate system of the camera reference frame. The z-axis of this coordinate system is perpendicular to the image plane.

Let X be the coordinate vector of point P in the world coordinate system. First, the 3x3 rotation matrix $\mathbf{Rc}$ is calculated from the 3-dimensional rotation vector ω by the Rodrigues Rotation Formula:

$$\mathbf{Rc} = \text{rodrigues}\ (\omega c)$$

The coordinate vector X_c in the camera reference frame is then given by the rigid motion equation:

$$X_c = \mathbf{Rc} \times X + Tc$$

By determining the extrinsic parameters for the left and right camera, the relationship between the two cameras can be derived.

Stereo parameters

The triangulation algorithm requires a mapping from coordinates in the left camera reference frame to the right camera reference frame. Given the rotation and translation vectors from the left camera ωc_L, Tc_L and from the right camera ωc_R, Tc_R, the transformation from a vector in the left camera reference frame to the right camera reference frame is given by

$$\omega s = \text{rodrigues}^{-1}\ (Rc_R \times Rc^T_L)$$
$$Ts = Tc_R - \text{rodrigues}\ (\omega s) \times Tc_L$$

with

$$Rc_L = \text{rodrigues}\ (\omega c_L),\ Rc_R = \text{rodrigues}\ (\omega c_R)$$

Since the two cameras are mounted at a fixed distance to each other, the parameters ωs and Ts can be determined once in the calibration process, before the stereo camera is installed at a specific location.

Calibration of intrinsic parameters

Due to the nonlinear nature of the described camera model, simultaneous estimation of the parameters involves applying an iterative algorithm. The Levenberg-Marquardt optimization method can be used to determine their values [6].

We currently use CalibFilter from Intel's Open Computer Vision Library to determine the intrinsic camera parameters, which provides a convenient solution to the calibration process. During calibration, a planar checkerboard is successively placed at various angles and distances with respect to the camera. The software automatically detects the reference points in the image and does not require any additional manual steps by the user.

Because the intrinsic parameters compensate for non-linear distortions, a large number of points in the reference pattern is required.

Calibration of extrinsic parameters

We use our own calibration software to determine the extrinsic and stereo parameters. The checkerboard turned out to be unsuitable, because the reference pattern has to be fully covered by both cameras at the same time.

When the checkerboard is placed too close before the cameras, it is not fully visible in both camera views at the same time. On the other hand, when the checkerboard is moved too far away from the cameras, the image resolution is too low to reliably detect all checkerboard fields. This effect is aggravated by the use of wide-angle lenses.

We use a board with five infrared emitting diodes as a reference pattern. Because the reference points emit light, they are easy to detect even over greater distances and nearly independent of the light conditions. The small number of reference points is sufficient because the extrinsic transformation is linear.

3. Position finding

The light emitted from an infrared diode on a tag appears as a bright spot in the image. The size and intensity of a spot depend on the distance of the tag to the camera. In the image processing step, these spots must be reliably detected and measured, ignoring other disturbing bright areas and noise.

The field of mathematical morphology [12] gives the theoretical foundations and building blocks for the necessary image processing operations.

1. In the first step, the grayscale image captured by the camera is converted to a binary image with a hysteresis threshold operator. This method proved to be quite robust against the considerable sampling noise of the camera.

2. Next, the algebraic area opening is calculated and subtracted from the image. This operation takes a pixel count as parameter and eliminates all bright areas larger than this threshold. Examples for such disturbing bright areas are lamps or sunlight on a window sill or table.

3. Now, the center coordinates of each bright spot are determined. This is done by calculating the mean value of all bright pixel coordinates that belong to a spot.

Depending on the environment, the infrared light emitted from tags can be reflected on lamp reflectors or other shiny surfaces. Such reflecting regions can be excluded from processing by defining a mask image.

3.1. Tracking

Because the transmission of an identification code spans multiple frames, the image points must first be tracked on the basis of nearest neighborhood. The result of the image processing step is a list L of 2D-coordinates.

For each element in L, the corresponding element in coordinate list L_{-1} of the previous image has to be identified. Intuitively, each element is associated with the previous element with the lowest Euclidean distance. Since emitters can go dark for up to two frames when a '0'-bit of an identifier is transmitted or may become temporarily occluded, the algorithm to create the mapping must search for a global optimum.

The association is solved by using graph algorithms for the maximum pairing problem. It is implemented by determining the maximum flow in a graph.

3.2. Removing camera distortions

The result of the image processing step are distorted pixel coordinate vectors. Each vector p_p must now be undistorted into a vector p_n. This process is called normalization. However, because of the high degree distortion model, there is no general algebraic expression for the inverse mapping of the described camera intrinsic parameter equations. We use the numerical implementation from the Camera Calibration Toolbox for Matlab [2].

3.3. Triangulation

The triangulation problem consists of determining the space coordinates $[x_c,y_c,z_c]$ of a point P in the camera reference frame from the left and right normalized image coordinates $p_{nL}=[x_{nL},y_{nL}]$ and $p_{nR}=[x_{nR},y_{nR}]$.

Let $\omega_L=[x_{nL},y_{nL},1]^T$ and $\omega_R=[x_{nR},y_{nR},1]^T$ be the coordinate vectors of perspective projections of P on the image

planes. To transform these two vectors into the same coordinate system, ω_L is transformed into the right camera reference frame:

$$\omega_L' = \text{rodrigues}(\omega_s) \times \omega_L + T_s$$

ω_L and ω_R can now be used to describe two lines in parametric form as follows:

$$l_L = \{T_s + \lambda_L \omega_L : \lambda_L\}$$
$$l_R = \{\lambda_R \omega_R : \lambda_R\}$$

The solution is the intersection point of the two lines l_L and l_R. But due to measurement errors they may not intersect, so the least squares method is used to minimize

$$|(T_s + \lambda_L \omega_L) - (\lambda_R \omega_R)|^2$$

The solution is

$$\begin{bmatrix} \lambda_L \\ \lambda_R \end{bmatrix} = (A^T A)^{-1} A^T T_s \quad \text{with} \quad A = [\omega_L \,|\, \omega_R]$$

4. Support for multiple tags

In order to distinguish different tags on the receiver's side, the tags send out identification codes.

The sampling rate f_r of the receiver is fixed by the camera frame rate, e.g. 30 frames per second. In the ideal case, the sender would use a signaling rate f_s of 30 Hz and therefore transmit with exactly the same rate as the receiver. But this would require perfectly synchronized clocks in the whole system and therefore also a bi-directional communications interface at the sender.

In our approach, the sender's signaling rate is set slightly lower than the receiver's sampling rate. Single bits will be observed duplicated at the receiver in regular intervals. The coding scheme employed must be resilient to these bit errors to correctly decode the transmitted data.

The example in Figure 1 shows the transmission of an identifier with f_r=30 Hz and f_s=24 Hz. The camera's exposure time is 1/125th of a second. From the ratio between f_r and f_s follows that every 4th transmitted bit is observed duplicated at the receiver.

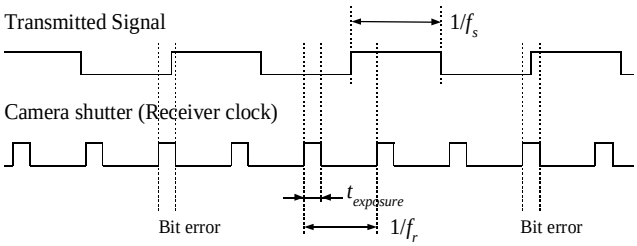

Figure 1. Transmission timing

The signaling rate must be higher than $f_r / 2$. The relationship between maximum signaling rate and camera frame rate is determined by the exposure time $t_{exposure}$. Two consecutive bit errors must be prevented. The maximum signaling rate is limited to

$$f_{smax} = 1/(\frac{1}{f_r} + t_{exposure})$$

Then the minimum number of consecutive bits that are transmitted correctly is

$$n_{min} = \frac{f_s}{f_r \bmod f_s}$$

4.1. Identifier encoding

For the design of a code to transmit identifiers, a trade-off between the following two concurring requirements must be found:

- **High tracking frequency:** The position of a tag can only be determined when the tag's emitter is on. This is only the case when the tag is sending out a '1'-bit (=light on). Therefore, the encoded signal should contain as few '0'-bits as possible.
- **High coding efficiency:** The longer the code words are and the more consecutive '0'-bits are permitted, the more efficient the coding scheme will be.

A simple variable length coding scheme that does not use repetitive '0'-bits is described in the following. The transmission of an identifier starts with the 'S' (start/stop) symbol (encoded as 111110). The identification number is then encoded by concatenating '0' and '1' symbols (encoded as 10 and 1110). At maximum, every 3rd bit may be duplicated in order to allow the receiver to correctly decode the transmitted identifier.

More efficient coding schemes are possible, if n_{min} is sufficiently high and/or more consecutive '0'-bits are permitted.

5. Implementation

The stereo camera consists of two off-the-shelf USB cameras mounted in a distance of 20cm on a DIN rail. Usually these cameras are targeted at video conferencing or similar applications and are equipped with lenses that have a relatively narrow angle. To increase the field of view, the cameras were fitted with 120° wide angle lenses.

These standard CCD-cameras also show a high sensitivity in the near infrared range and register the infrared diodes transmitting at 880nm as bright spots. An infrared filter was added to increase the contrast between IR and visible light that ranges from about 800nm to 400nm. This filter is improvised of one or two layers of unexposed but developed E6 film. Two layers are basically visually opaque with transmission increasing

rapidly starting at 720nm to about 90% at 850nm. Further, the image sharpness is not degraded significantly. This interesting option to professional photographic filters was suggested in [5], where the author investigated E6 film with a spectrophotometer.

The cameras are connected to a PC with an USB 2.0 host controller or two different USB 1.1 host controllers, as each camera consumes one full USB 1.1 bandwidth. Because the software does extensive real-time image processing and floating point operations, a 2.4 GHz Pentium-4 or an equivalent dual-processor system is required for two cameras at 640x480 and 30 frames per second. Because of USB bandwidth limitations, the captured images are compressed in the camera and have to be decompressed by the camera driver in software. This decompression consumes about 50% of the total processing time. Increasing the number of tags (up to about 100) does not have a significant impact on processing time.

The running system consists of three components. Two instances of the ActiveMovie-based filter SpotTracker process the left and right camera images, respectively. The service component merges the information from the two filters, calculates the 3D-coordinates and exposes the measured location information to applications via a XML/SOAP-based interface.

These three components communicate via shared memory. Beside providing image coordinates, the SpotTracker filter also generates an output image for inspection purposes. This image consists of the input image from the camera with the detected spot coordinates and decoded identifiers overlaid. Figure 2 shows a test setup with normal camera lens and removed IR-filter. The small tag in the lower right corner sends the identifier 10, the programmable LEDs on the Atmel developer board the identifiers 1-8. The three continuously active status LEDs do not emit identifiers and are therefore assigned negative numbers by the software.

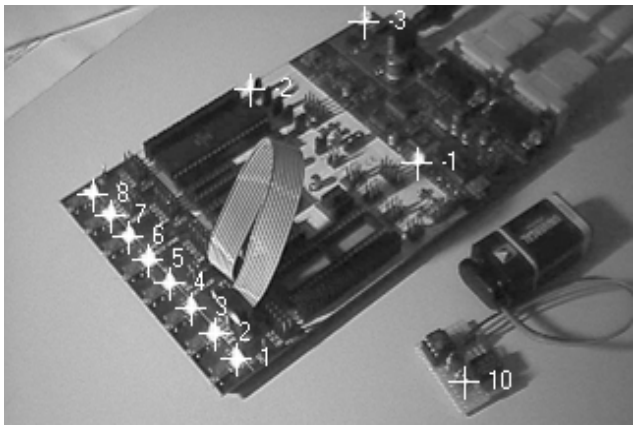

Figure 2. SpotTracker output image

6. Evaluation

The system was installed in a lecture hall to determine its accuracy. The camera was mounted in front of the blackboard at a height of about 3 meters. Because the camera is fitted with wide angle lenses it can cover nearly the full room which measures 15.1 x 9 meters.

The lecture hall is in the basement, has no windows and is illuminated by fluorescent light. All lights were turned on during the test. The infrared emitter consisted of one infrared LED with a narrow angle of ±20 degrees. One such LED easily ranges 10 meters and more. An exposure time of 1/500th second was sufficient to pick up the signals at the cameras. The camera resolution was set to 640x480.

The infrared emitter was then placed at one test point after the other. At each point, the position calculated by the positioning system was compared with the expected position and the error distance was calculated. We determined the expected positions by hand using a measuring stick. Thus, the accuracy of the reference system is limited.

The result is displayed in Figure 3. The Root Mean Square (RMS) error calculated from all 138 test points is 16.67 cm. The ±60° line shows the RMS error depending on the distance from the camera. The ±45° line shows the error for that subset of points that lies within a ±45° frustum and so on. We can see that the measurement accuracy decreases with increasing angle from the camera axis and with distance.

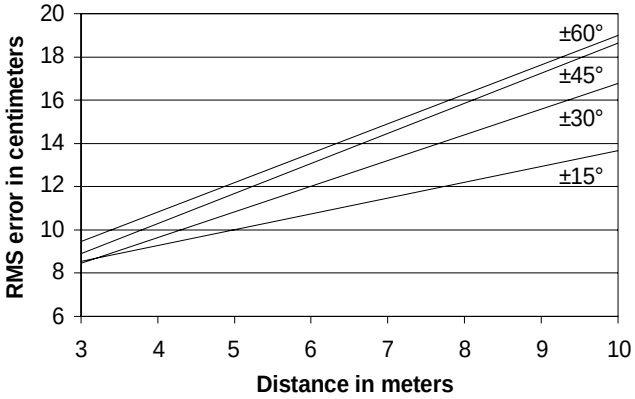

Figure 3. System accuracy

The current system can be improved in a number of ways. First, a higher camera resolution would be desirable. Our cameras have a 1.2 megapixel sensor, but this resolution can only be used in still mode. In streaming mode, only 640x480 pixels and 30 frames per second are supported due to the narrow USB 1.1 bandwidth constraints. Furthermore, the image compression algorithm employed by the camera seems to degrade image quality and therefore lowers the accuracy of the system. A higher image resolution would enable higher positioning accuracy. Identifiers could also be transmitted

faster with a higher frame rate.

The two USB cameras are not synchronized and therefore do not take pictures at exactly the same time. This means that fast motions will add an additional error to the position measurement. It is possible to reduce this effect by using interpolation.

6.1. Applications

The Talking Assistant [1] (TA) is a voice-centric user terminal in the form factor of a headset. It supports audio in-/output and features a hardware MP3-decoder for reasonable playback quality. The headset connects wirelessly to the network infrastructure via Bluetooth.

The device is location-aware by utilizing the infrared positioning system presented in this paper. Eight IR emitter diodes are mounted on top of the headset, each covering an angle of 45 degrees. We use this large number of emitters with a relatively narrow angle to achieve a greater range of the system. Only the emitter that is oriented towards the receiving stereo camera is used to send out a beacon signal at any time. The direction of the stereo camera is given by the heading information, the current position and the world model.

Because the emitters are worn on the head, they are visible most of the time to a ceiling-mounted camera. In this case, the direct line of sight requirement is hardly a restriction.

A combination of two sensors is used to determine the user's head orientation in all three rotational axes. An acceleration sensor is used to measure the tilt angles and the heading is measured with an electronic compass.

The TA headset targets tour, exhibition, and museum guidance as an initial application domain. The following list contains key requirements we identified for end-devices used to augment the experience of exhibitions:

- **Eyes free:** Textual explanations on signs or displays distract the user from examining the exhibits, thus reducing the overall experience of the exhibition. With TAs, visitors can naturally explore the environment.
- **Hands free:** Following the invisible computing paradigm, the technology should act unobtrusively in the background. This requires small hardware that does not burden the user and is easy to carry.
- **Location aware:** The location system must be accurate enough to determine which exhibit the user is currently looking at. This requires an accuracy in the centimeter range and measuring of the user's head orientation.
- **Personalized:** Personalization (e.g., reflecting the user's interests, available time, and tour history) is the single most important advantage over conventional electronic guides.
- **Networked:** The wireless network interface permits the device to act as a “thin client” that relies on services in the network. The computational power is in the service infrastructure. This avoids complexity at the end-device.

In a lecture scenario, the Talking Assistant device can serve as a radio microphone. When recording the lecture on video, the cameras can be automatically reoriented onto the lecturer with the position information provided by the location system.

For such applications, the positioning cameras can be placed opposite to the recording camera. Thus, no beacon signals are sent towards the video camera. These would result in disturbing bright spots on the recording, because virtually all cameras are also sensitive in the infrared spectrum.

7. Conclusion

We have built and evaluated an infrared optical local positioning system for research and applications in ubiquitous computing. It offers an accuracy of about 8cm in near range and about 16cm when covering ~100m^2. In future work, we plan to increase the accuracy of the system and to investigate hybrid systems by adding an inertial tracker. This relative positioning sensor could bridge times when no direct line of sight is available.

References

[1] E. Aitenbichler, and M. Mühlhäuser, The Talking Assistant Headset: A Novel Terminal for Ubiquitous Computing. Technical Report TK-02/02, TU Darmstadt, 2002.

[2] J. Y. Bouguet, Camera Calibration Toolbox for Matlab, http://www.vision.caltech.edu/bouguetj/calib_doc/

[3] J. Caffery, and G. Stuber, Subscriber Location in CDMA Cellular Networks. IEEE Transactions on Vehicular Technology, Volume 47 Number 2, 1998, pp. 406-416.

[4] Cybernet Interactive, Firefly Motion Capture System, http://www.cybernet.com/interactive/firefly/

[5] A. Davidhazy, Making an Improvised Infrared Transmitting Filter, http://www.rit.edu/~andpph/

[6] J. Heikkilä, and O. Silvén, A Four-step Camera Calibration Procedure with Implicit Image Correction. CVPR 1997, pp. 1106-1112.

[7] J. Hightower and G. Borriello, Location Systems for Ubiquitous Computing. Computer, Issue on Location Aware Computing, Volume 34 Number 8, 2001, pp. 57-66.

[8] J. Latvala, J. Syrjärinne, S. Niemi and J. Niittylahti, Patient Tracking in a Hospital Environment Using Extended Kalman-filtering. Proc. IEEE Middle East Workshop on Networking, Beirut, Lebanon, 1999.

[9] Northern Digital Inc., Optotrak, http://www.ndigital.com/

[10] K. Pahlavan, X. Li and J. P. Mäkelä, Indoor Geolocation Science and Technology. IEEE Communications, February, 2002, pp. 112-118.

[11] RF-Technologies, PinPoint Asset Tracking Solutions, http://www.rftechnologies.com/pinpoint/

[12] P. Soille, Morphological Image Analysis. Springer-Verlag, 1999.