Sensor Node Localisation using a Stereo Camera Rig

Dermot Diamond, Noel E. O'Connor, Alan F. Smeaton,
Stephen Beirne, Brian Corcoran, Philip Kelly, Kim Tong Lau, Roderick Shepherd,
Centre for Digital Video Processing & Adaptive Information Cluster,
Adaptive Information Cluster, Dublin City University, Ireland
{Dermot.Diamond, Noel.OConnor, Alan.Smeaton}@dcu.ie

ABSTRACT

In this paper, we use stereo vision processing techniques to detect and localise sensors used for monitoring *simulated environmental events* within an experimental sensor network testbed. Our sensor nodes communicate to the camera through patterns emitted by light emitting diodes (LEDs). Ultimately, we envisage the use of very low-cost, low-power, compact microcontroller-based sensing nodes that employ LED communication rather than power hungry RF to transmit data that is gathered via existing CCTV infrastructure. To facilitate our research, we have constructed a controlled environment where nodes and cameras can be deployed and potentially hazardous chemical or physical plumes can be introduced to simulate environmental pollution events in a controlled manner. In this paper we show how 3D spatial localisation of sensors becomes a straightforward task when a stereo camera rig is used rather than a more usual 2D CCTV camera.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*depth cues, stereo*

Keywords

Sensor Nodes, Localisation, Stereo

1. INTRODUCTION

In previous work, we showed how LED communications from sensor nodes can be picked up in a controlled environment using a simple CCTV camera [3]. Chemical gas plumes were monitored by low-cost LED-based chemosensors and when threshold chemical concentrations were detected, a signal LED was used to communicate this event to the camera. We have also reported on the use of conventional RF-enabled nodes with the same LED-based sensors [13]. Looking to the future, and as also recognised by others (see [5], for example), we feel that the most realistic approach to deploying over wider areas will require a combination of these data harvesting approaches, with specific configuration dictated by the application in question. In this paper, we are again interested in understanding the potential of using cameras to harvest sensor data, however this time we consider a more sophisticated camera device as the data gatherer in order to facilitate automatic node localisation.

EmNets'07, June 25-26, 2007, Cork, Ireland.

2. CCTV & STEREO VISION

In our initial proof of concept work, we postulated that sensor nodes could be placed within a camera's field of view and that the sensors could signal their status (and what they are sensing) to the camera via LEDs, thereby avoiding power-hungry RF transmissions. Clearly, for such an approach to exhibit practical scalability to real scenarios, sufficiently dense CCTV coverage is required and some means to automatically localise arbitrarily placed sensor must be available.

Such visual sensing coverage is already in place within most built environments such as city centres, private property and strategic buildings where people congregate (including subways, train stations, airports, etc.) via the growing deployment of CCTV [11]. Using the existing CCTV infrastructure would mean that battery-life concerns associated with visual processing on power-constrained devices can be side-stepped since the cameras are wired and support continuous sensing. In fact, it even becomes feasible to start considering a change to the camera modality and/or the possibility of pushing relatively computationally expensive data processing to the camera.

Existing CCTV cameras simply gather and transmit the video data to a back-end server. Thus, the gathered data is really only useful in a forensic capacity i.e. after an important event has occurred. Ideally, we would like to locate the required image processing in the camera itself so that it can 'wake-up' upon detection of an event, and signal this appropriately. Of course, this would necessitate replacing every camera in the network with a 'smart' camera with the required processing capability. However, CCTV networks will be upgraded in time and indeed replaced as they get older. Whilst the camera placement is expensive, it is a once-off activity that can be thereafter used over a lengthy period for multiple deployments of cheap sensors.

Our initial work used 2D visual processing for sensor detection but did not consider node localisation. One approach to address localisation is outlined in [7], whereby each mote

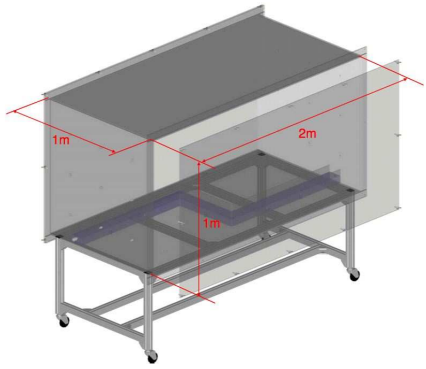
is equipped with a LED and is wireless enabled. Computer vision techniques are employed to obtain the direction of the continuously blinking LEDs from a single camera. The distance of the mote is approximated by obtaining the power drop in RF signal strength. However, this solution employs idealised models to infer distance from received signal strength, which may be violated in some scenarios. In addition, the use of wireless has implications on the power requirements of the motes. An interesting discussion on the variation of received RF signal strength due to diffraction, refraction, reflection and multi path fading as well as RF power issues is discussed in [10].

A particularly elegant approach, outlined in [2], eliminates the need for wireless communication between non-visual sensor motes by employing only computer vision techniques to establish node localisation in the shared field of view of a number of widely distributed camera sensor platforms. The approach employs the use of modulated light emissions from a bright red LED as both a communication and localisation mechanism. This reduces the cost of deployment since individual nodes will not be required to create and maintain their own communications infrastructure and so they can be built as simpler, cheaper low power devices. In the approach proposed in [2], it is necessary that each LED is able to uniquely identify itself to the cameras and this is achieved by modulating the LED emissions from each mote. Once each camera has identified each of the motes present in its field of view, the information from multiple cameras is merged to calibrate the cameras and thereby localise the motes.

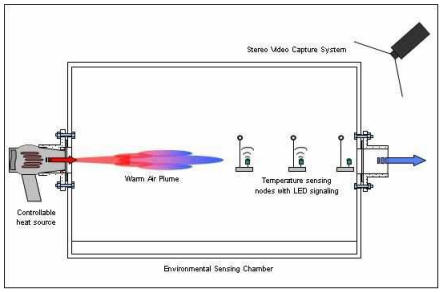
In this work, we consider an alternative approach based on the idea of using existing wired CCTV infrastructure, albeit an infrastructure where the 2D cameras are replaced with narrow base-line stereo cameras as a result of our hypothetical upgrade mentioned above. Such devices are likely to become widespread in future CCTV networks, given the benefits they have been proven to provide over 2D cameras in terms of robust analysis of unconstrained scenes for security and safety applications – see [6, 9]. The use of calibrated short-baseline stereo cameras eliminates the need for each mote to uniquely identify itself in addition to avoiding the requirement of [2] for the deployment of special patterns that includes distance information, or distance measurements between camera nodes. Notwithstanding this, we acknowledge that a stereo camera network will still suffer the limitations of any such approach (e.g. line-of-sight issues between nodes and cameras, temporary occlusions, shadows, etc.), and thus our future work will consider a similar approach to that of [2] but with multiple stereo cameras with overlapping fields of view.

3. ENVIRONMENTAL CHAMBER

To facilitate our work on these and other issues, we have built a laboratory-based *environmental sensing chamber* (see figure 1(a)) where we can locate LED-based sensor nodes and introduce gas plumes in a controlled manner in order to simulate an environmental pollution event – see figure 1(b). The chamber is constructed from bonded 10mm thick clear Perspex^{TM} sheets and has a dimension of $2m \times 1m \times 1m$. This constitutes a versatile airtight environment that allows for the use of a range of hazardous chemicals and other mediums during testing. We use Mica2Dot MPR500CA mote devices [1]. Since the work reported here is a proof of concept



(a)



(b)

Figure 1: Environmental Sensing Chamber: (a) Dimensions; (b) System diagram showing air flow, sensors and stereo camera. Note that in practice sensors are placed in arbitrary locations in the 3D volume.

of using a stereo camera, for our experiments in this paper we simply sense temperature rather than any hazardous chemical. It should be noted though that we have a pipeline of other sensing modalities we are working with, including very low cost optical chemo-sensors that can be used as detectors in gas and liquid phase monitoring applications [14, 12]. We employ a single thermistor per mote interfaced to a single ended analog to digital converter (ADC). ADC values are used unconverted to trigger a signalling LED when a preset threshold temperature is reached. An example of a sensor node is shown in figure 2(a).

4. VIDEO PROCESSING

4.1 Pre-processing

Node localisation is achieved via a triangulation process based upon establishing correspondence between the same scene points imaged by the rig's two cameras – see figure 3. The difference from where the point u occurs in the left image and where the corresponding point u' occurs in the right image is known as the *disparity*. Many disparity estimation techniques simplify this correspondence problem by applying a well known geometric rule known as the *Epipolar Constraint* [15]. By applying this rule, the correspondence problem is reduced from a 2D to a 1D search along the so-called *Epipolar Line*. To obtain the disparity of the LEDs, we employ a *correlation based* disparity estimation technique [4] that assumes that corresponding pixels between images

(a) (b)

Figure 2: Sensor nodes: (a) An example sensor node incorporating a Crossbow Mote (right) and LED sensing/signalling (left); (b) Thermistor potential divider circuit

have very similarly intensities. We therefore perform image colour normalisation prior to this. Following this, the images from the two cameras are *rectified* i.e. they are projected onto a new common image plane that is parallel to the cameras' baseline [4].

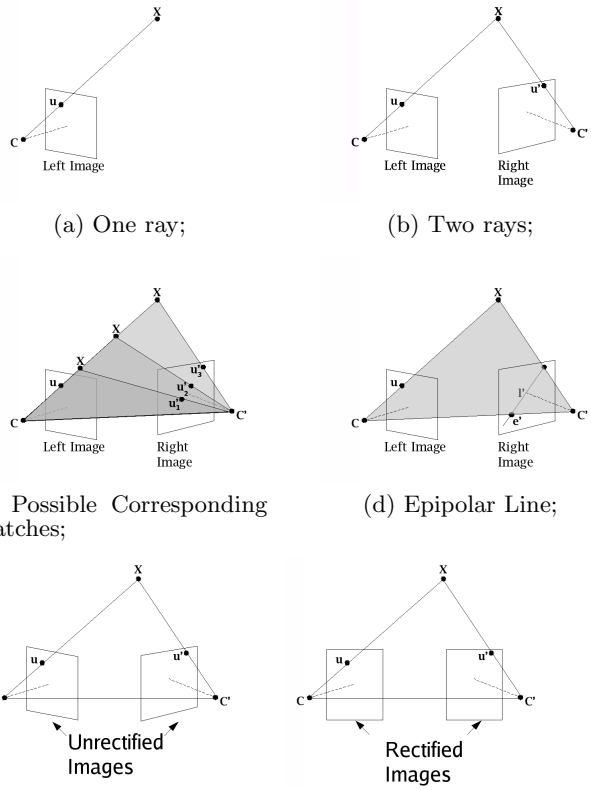
4.2 Foreground Extraction

Prior to performing disparity estimation, it is necessary to extract the foreground LED regions. A commonly used technique is *background suppression* – see [16, 8]. This involves building a representation of the background and then detecting foreground pixels as those that differ significantly from their modeled value. However, unlike most classical background suppression scenarios that attempt to solve a two-class classification problem where the foreground is undefined, in our scenario, we know the colour of the LEDs we are attempting to detect. This allows us to create a robust background suppression algorithm that is less sensitive to noise. Our algorithm is implemented in two steps applied to both images separately:

- Background Suppression: We employ a Gaussian based background modeling technique, similar to that defined in [17], to obtain foreground regions.
- Foreground Colour Mask: The foreground regions are then filtered through a colour mask, which extracts green coloured foreground regions (the LED colour we use). This is a natural extension of the 2D image processing we used in [3].

4.3 Disparity Estimation

In order to obtain the 3D position of each detected region, corresponding to the spatial location of the LED, the disparity of each foreground region must be determined. An initial step to this process is to cluster the pixels in one foreground image, FI_1, into regions using a 4-neighbourhood connected component algorithm [4]. For a given foreground region, r_1, in FI_1, the disparity of the region is obtained by moving r_1 along its corresponding epipolar line for a maximum distance of d_{max}, the disparity limit. At each position along the epipolar line, a matching score is obtained using the *Sum of Absolute Differences* (SAD) between all the pixels in r_1 and the corresponding pixels in FI_2. The position along the epipolar line where the matching score is minimum is chosen as the corresponding region, r_2, iff r_2 contains at least 1



(a) One ray; (b) Two rays;

(c) Possible Corresponding Matches; (d) Epipolar Line;

(e) Unrectified Images; (f) Rectified Images;

Figure 3: (a) & (b) Triangulation; (c) & (d) Epipolar Geometry; (e) & (f) Rectification

foreground pixel. The disparity is then defined as the difference in position of r_1 and r_2 within their respective images. This process is then reversed to obtain the disparities for all regions in FI_2 to remove any inconsistent region matches.

4.4 3D localisation

The chamber is positioned in 3D space relative to the camera by manually tagging 3 points in the scene in both I_1 and I_2. This is carried out once, when the camera is initially positioned (the once-off calibration step required at installation time, referred to in section 2). The disparity of these points are determined and the 3D position of the points is calculated from the rectified images using triangulation, as defined in [15]. These 3 points are used to constrain the chamber's position and orientation in 3D space with respect to the camera. Each *valid* foreground region, as obtained from section 4.3, is also projected into 3D space. It is then possible to obtain the 3D coordinates of the projected region with respect to the environmental chamber's coordinate system, in fact with respect to the entire scene.

5. EXPERIMENTAL RESULTS

Experimental data was captured using a Digiclops© stereo camera at a distance of 1.4 meters to the chamber. The camera baseline is 10cm and image resolution is 640×480 pixels. Four separate tests were carried out, resulting in 2470 images, where upto 3 LEDs may be on at any given

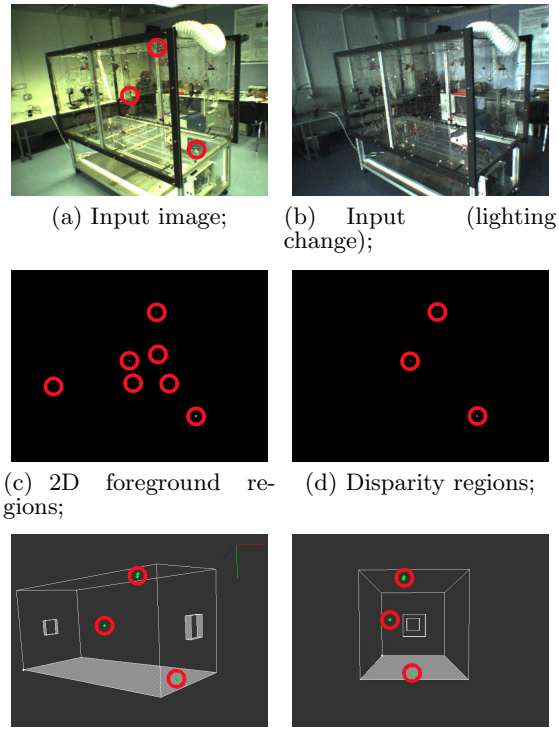
time. The experiments involved various mote setups, where the motes were positioned at known ground truth positions with respect to the chamber's coordinate system. In addition, different lighting conditions were simulated. Each mote is equipped with a green LED. An LED is determined as *detected* if the foreground region in I_1 and its corresponding region in I_2, both correspond to a real LED being on within the chamber.

To evaluate the detection results we use the commonly used metrics of *Precision (P)* and *Recall (R)* defined as:

$$P = \frac{LED_{TP}}{LED_{TP} + LED_{FP}} \qquad R = \frac{LED_{TP}}{LED_{TP} + LED_{FN}}$$

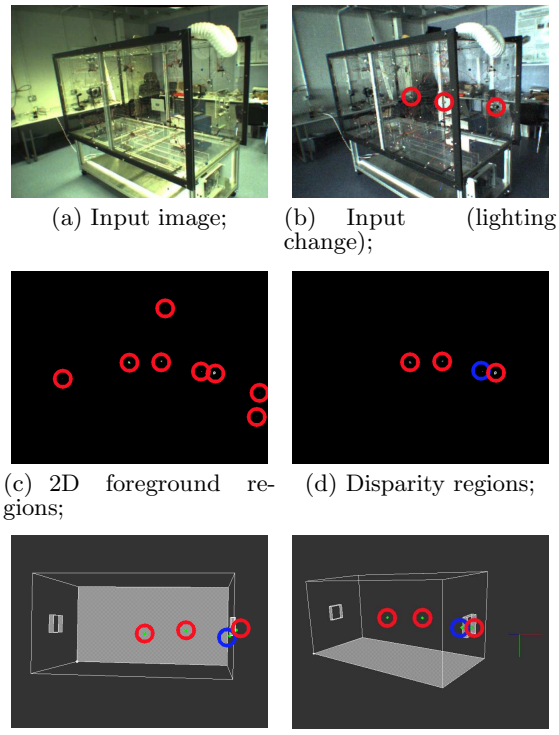
where LED_{TP} is the number of true-positive LED regions, i.e. the number of LED regions correctly detected by our system, LED_{FP} is the number of false-positive LED regions, i.e. the number of incorrect LED detections, and LED_{FN} is the number of false-positive LED regions, i.e. the number of missed LED detections. For a total of 3,658 LED activations, we obtained an average precision of **95.35%** and an average recall rate of **99.18%**. The average difference between the *real-world* position of an LED and the *detected* position of an LED that has been switched on, with respect to the chamber's coordinate system was **15.63cm**, where the average distance from an LED to the camera was 215.37cm. Thus, we obtained the correct 3D position of LEDs with an error in precision of just **7.258%** with respect to the camera placement. These results indicate extremely accurate detection and spatial localisation. Many of the false-positives in precision are due to reflections of activated LEDs from the perspex sides that occur when an LED is positioned at the edge of the chamber.

Figures 4 and 5 show illustrative results for two different spatial configurations of 3 sensors within the chamber. Figures 4(a) and 5(a) show the colour images obtained by one of the camera lenses, whilst figures 4(b) and 5(b) show the colour images obtained using the same experimental set-up in each case but under different lighting conditions. The ground truth position of the motes for each configuration are shown using red circles in one of the images. For both spatial configurations, typical results of node detection using 2D image processing (such as that used in our previous work) is illustrated in figures 4(c) and 5(c), where detected nodes are again indicated by red circles. From these images it is clear that there are many erroneous detections that cannot be resolved with further 2D post-processing. Detection results from the stereo rig are depicted in figures 4(d) and 5(d) indicating much improved detection accuracy. Figure 5(d) illustrates a problem with the approach whereby it can be seen that one of the "detected" motes is a false-positive that results from the reflection of the LED on the surface of the perspex side of the chamber – this is illustrated by a blue circle. In figures 4(e) and 4(f), the LEDs are shown rendered as 3D points in the correct spatial location in a stylised wireframe model of the chamber from two different artificially generated viewpoints (where the shaded panel indicates the "floor" of the chamber), and similarly in 5(e) and 5(f) for the second spatial configuration. Such a rendering of the environment being sensed could feasibly be transmitted/stored as a low bandwidth alternative (or indeed as a complement) to the video data itself.



(a) Input image; (b) Input (lighting change); (c) 2D foreground regions; (d) Disparity regions; (e) 3D Positioning – View 1; (f) 3D Positioning – View 2.

Figure 4: Illustration detection results 1



(a) Input image; (b) Input (lighting change); (c) 2D foreground regions; (d) Disparity regions; (e) 3D Positioning – View 1 (f) 3D Positioning – View 2;

Figure 5: Illustrative detection results 2

6. CONCLUSIONS

These experiments are extremely encouraging and demonstrate the potential usefulness of stereo vision as a means to harvest data from light emitting sensor networks. This serves to convince us of the possibilities of leveraging already instantiated CCTV infrastructure as a data gathering framework for wireless sensor networks, albeit necessitating increased sophistication in the camera device itself. We have proved that using mature well understood and computationally efficient vision techniques it is possible to detect and spatially localise the sensors within an environment, facilitating more accurate detection and monitoring of an event as it unfolds. Clearly, this approach is inherently limited by the distance of the sensor node to the camera as the brightness of the LED will diminish significantly the further away it is, making detection problematic and error prone. In this work to ignore this, as the next step in our work is to use multiple stereos with overlapping fields of view under the hypothesis in such a scenario, an LED is always sufficiently close to one camera in the network to facilitate detection.

We acknowledge that whilst the use of the environmental chamber in this paper is useful from a conceptual point view, and provides an easily controlled testbed for the next phase of our work with more complex sensors with real chemical events, it is also somewhat artificial. For this reason, we have to date deliberately not tackled vision problems introduced by the chamber itself. An example is the reflections caused by the perspex, that could be easily removed from consideration by determining that they are outside the 3D volume of the chamber.

In the future, we plan to use the optical chemo-sensors mentioned in section 3. In parallel we will also distribute a larger number of sensors with more advanced light signalling in various 3D spatial configurations within a real environment, corresponding to a large room or corridor as a first step, with sensors artificially activated based on modeled dispersion characteristics. We will also move to using multiple stereo rigs trained on the same scene and use this to disambiguate occlusions, motivated by the approach of [2]. We will perform experiments in active scenes i.e. with moving people/objects, using the stereo techniques we have developed for foreground object detection and tracking in surveillance scenarios.

Acknowledgments

This work is supported by Science Foundation Ireland under grant 03/IN.3/I361.

7. REFERENCES

[1] Mica2dot wireless microsensor mote document part number: 6020-0043-05 rev a. 2005., 2005.

[2] A. Barton-Sweeney, D. Lymberopoulos, and A. Savvides. Sensor localization and camera calibration in distributed camera sensor networks. In *Proceedings of IEEE BaseNets*, 2006.

[3] E. Cooke, N. O'Connor, A. Smeaton, D. Diamond, R. Shepherd, S. Beirne, and B. Corcoran. Video analysis of events within chemical sensor networks. In *ICOB 2005 - 2nd Workshop on Immersive Communication and Broadcast Systems*, October 2005.

[4] D. Forsythe and J. Ponce. *Computer Vision - A Modern Approach*. Prentice Hall, 2003.

[5] P. B. Gibbons, B. Karp, Y. Ke, S. Nath, and S. Seshan. Irisnet: An architecture for a worldwide sensor web. *IEEE Pervasive Computing*, 2(4), 2003.

[6] M. Harville. Stereo person tracking with adaptive plan-view templates of height and occupancy statistics. *Image and Vision Computing*, 22(2):127–142, 2004.

[7] S. Hengstler and H. Aghajan. Application development in vision-enabled wireless sensor networks. In *Proceedings of Systems and Networks Communications (ICSNC)*, 2006.

[8] P. KaewTraKulPong and R. Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. In *2nd European Workshop on Advanced Video-based Surveillance Systems, Kingston upon Thames*, 2001.

[9] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer. Multi-camera multi-person tracking for easyliving. In *Proceedings of IEEE Workshop on Visual Surveillance*, 2000.

[10] P. Natara, T. N. Mundhenk, K. Bellman, M. A. Arbib, and L. Itti. Camera localization methods for intelligent room systems using RF techniques. In *SPIE Conference on Intelligent Robots and Computer Vision XXII: Algorithms, Techniques, and Active Vision*, volume 5608, pages 177–187, 2004.

[11] C. Norris, M. McCahill, and D. Wood. The growth of CCTV: a global perspective on the international diffusion of video surveillance in publicly accessible space. *The Journal of Surveillance and Society*, 2 Summer:110–135, 2004.

[12] M. OToole, K.-T. Lau, B. Shazmann, R. Shepherd, P. N. Nesterenko, B. Paull, and D. Diamond. Novel integrated paired emitter-detector diode (PEDD) as a miniaturized photometric detector in hplc. *Analyst*, (131):938943, 2006.

[13] R. Shepherd, S. Beirne, K. Lau, B. Corcoran, and D. Diamond. Monitoring chemical plumes in an environmental chamber with a wireless chemical sensor network. *Sensors and Actuators B*, 121:142–149, 2007.

[14] R. L. Shepherd, W. Yerazunis, K. Lau, and D. Diamond. Low-cost surface-mount LED gas sensor. *IEEE Sensors Journal*, 6(4):861–866, Aug 2006.

[15] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis and Machine Vision, Second Edition*. PWS Publishing, 1999.

[16] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings of CVPR99*, pages II:246–252, 1999.

[17] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder. real-time tracking of the human body. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 19, pages 780–785, 1997.