Multimodal abandoned/removed object detection for low power video surveillance systems

Michele Magno, Federico Tombari, Davide Brunelli, Luigi Di Stefano and Luca Benini
DEIS - Universitá di Bologna
Viale Risorgimento 2, 40136 Bologna, Italy
{name.surname}@unibo.it

Abstract

Low-cost and low-power video surveillance systems based on networks of wireless video sensors will enter soon the marketplace with the promise of flexibility, quick deployment and providing accurate and real-time visual data. Energy autonomy and efficiency of the implemented algorithms are undoubtedly the primary design challenges to be addressed on systems subject to low computational capabilities and memory constraints. In this paper we present a low-power video sensor node designed for low-cost video surveillance which is able to detect abandoned and removed objects. The system exploits multimodal sensor integration which saves on-board power consumption. In particular a Pyroelectric InfraRed (PIR) sensor is exploited to optimize the use of the camera, grabbing images only when required in order to obtain the maximum efficiency from event recognition. Our fixed-point ARM-based approach is characterized in terms of runtime execution and power consumption, while efficiency is demonstrated by experimental results and compared with floating point implementations.

I. Introduction and previous work

The demand for reliable surveillance systems is increasing, especially for mass transit and public areas such as airports, railway and subway stations, sport and concert event venues. For this reason, video surveillance systems that, through the analysis of video sequences, perform automatic detection of security-related events or aid human personnel at monitoring a place are gaining increasing interest. A key aspect for current video surveillance systems is the capability of reliably detecting common events such as abandoned and removed object within the scene. Typical scenarios are, e.g., detection of unattended packages in a railway station or in an airport [1], [2], and detection of stolen objects in a museum [3]. Nevertheless many proposals have recently addressed this specific task [1]–[12], none of them are based on an embedded and unobtrusive architecture able to be long-term operating, to execute surveillance algorithms completely locally and to rise alarms wirelessly only when suspicious events happen.

We aim at filling this gap by proposing a multi-modal video surveillance system, characterized by low power consumption and low cost, and based on a CMOS video sensor and a Pyroelectric InfraRed (PIR) sensor. The use of the PIR sensor can notably reduce the overall power consumption of the system in absence of events, as shown in [13], where an embedded video system has been designed to detect structural effectively and rapidly changes in the monitored scene by jointly exploiting camera and PIR. The objective of this work is to propose a more advanced video analysis framework that, based on similar low-cost and low-power architecture, is able to detect events such as abandoned or removed objects.

Recently, applications which exploit Low-power Video Wireless Networks (LP-VWN) consisting of networks of low-cost video sensors connected by low-rate wireless channels and constrained by low-power budget, have gained increasing attention. LP-VWNs, in fact, represent a strategic enabling technology for a number of applications in surveillance, environmental monitoring, entertainment and health care. Designing a distributed video system within the tight power budget typical of mobile devices and wireless sensor networks is a very challenging task. Typical applications are in the domain of object detection or tracking.

When an event is detected, if the full image is not essential for the particular application, the system may transmit only some very limited amount of information, such as number of objects, size, position, trajectory, etc. saving a large amount of energy in wireless transmission

978-0-7695-3718-4/09 $25.00 © 2009 IEEE
DOI 10.1109/AVSS.2009.72

and extending the autonomy of the batteries. Clearly, nothing should be done from the point of view of data transmission and power consumption if the targeted object is not detected because simple raw cameras are exploited. In this case, the detection of abandoned or removed objects can be performed only after the collection of continuous video streams transmitted to cumbersome power-unconstrained base station. Of course this approach would be extremely energy and bandwidth inefficient, difficult to port on stand-alone mobile embedded systems and ultimately not scalable in a network. Smart wireless video networks architectures are possible only if they are based on devices with an adequate trade-off between power consumption and processing capabilities, thus the key challenges we addressed are the development of energy-efficient algorithms and low-power architectures which can support vision-related tasks.

Research on low-cost video node design has been very active in the last years and a number of node prototypes have been designed [14]–[20]. We can classify these approaches in three categories: (i) low-cost nodes with wired interface (e.g., the node designed by Corely et al. at CMU [15]), (ii) wireless nodes with significant power consumption (e.g., the Panoptes nodes designed by Feng et al. [18]), (iii) application specific single ultra-low power single chip solution (e.g., the chip designed by Zhang et al. [17]). Nodes in the first category obviously do not satisfy the basic requirement of being wireless, while nodes in the second category consume roughly 10x more power than typical nodes in a wireless sensor networks. Finally, the single-chip solution have extremely low power consumption, but it is not programmable nor configurable in field. One important common point in current video wireless nodes of the first and second category is that the digital signal processing subsystem is the main power bottleneck. This is due to the fact that the high data rate of CMOS image sensor imposes the selection of fast processors and memories with high power consumption. Hence, the main open challenge in this area is to synergically develop algorithms and architectures for energy-efficient image processing without giving up the flexibility of in-field configuration.

Energy autonomy and efficiency of the implemented algorithms are undoubtedly the primary design challenges to be addressed on systems subject to low computational capabilities and memory constraints. Both issues are addressed by the integration of multi-modal information using additional ultra-low power PIR sensors which increases energy efficiency because the camera is triggered only when necessary and, in the same time, reduce considerably the average power consumption of the wireless video node because camera is in shutdown state in absence of events.

Other work presented a combination of video sensor with other low-cost and low-level sensors, which are used mainly for triggering the camera at the right time and not to promote a reduction of the system energy requirements. A distributed network of motes equipped with PIR, acoustic and magnetic sensors with adjustable sensitivity have been proposed in [21], stealthiness and effectiveness in a military surveillance applications. A network of IR sensors and cameras are used also in [22] to balance privacy and security in surveillance applications.

We present a video sensor architecture designed for low-power and low-cost video surveillance centered around a STR912F from ST-Microelectronics equipped with an ARM966E 16/32-bit RISC, 96 MHz operating frequency, 96 KB SRAM and several interfaces. We implemented an algorithm for detecting abandoned and removed objects within the scene which is optimized for low-power architectures constrained by limited computational capabilities. The main constraints when developing algorithms for such architectures characterized by small available memory is efficiency and timing performance. Furthermore optimizations have to be implemented taking into account that a floating-point unit is unavailable. However, experimental results demonstrate the quality of our multi-modal ARM-based approach. Moreover we analyze different configurations and characterize the system in terms of runtime execution and power consumption, comparing the results of efficiency with floating point implementations on personal computers.

The remainder of the paper is organized as follows. In the next section we present the system architecture focusing on the constraints of energy budget, memory and computational capability offered by an ARM-based solution. The developed system and the description of the several power modes used by the application is also discussed.Section III depicts the algorithm implemented for the detection of abandoned/removed objects. In particular we discuss constraints and requirements of implementation on limited platform when optimizations are necessary. Experimental measurements and achieved performance are the focus of Section IV. Finally, Section V draws conclusions.

II. System architecture

The developed smart camera is showed in Fig. 2 and it consists of three modules: an multi-sensor layer (MSL) equipped with an image sensor and a pyroelectric sensor, a processing unit(PU) based on ARM9 architecture, and a wireless communication unit (WCU), as shown in Fig. 1.

The MSL includes a small PCB with 1 megapixel color CMOS imager VS6624.It supports up to 15 fps SXGA with progressive scan and up to 30 fps with VGA format with a typical power consumption of $120\,mW$ when active, while it decreases down to $23\,mW$ in stand-by mode. The

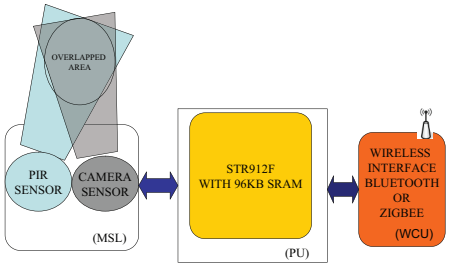

Fig. 1. Video sensor node architecture.

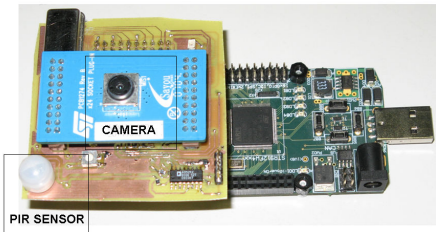

Fig. 2. Developed prototype of the video sensor node.

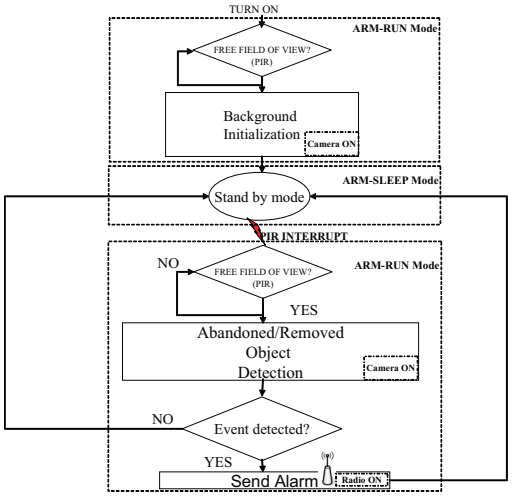

Fig. 3. Flow chart of the application.

system exploits PIR Sensor typically used in surveillance to provide simple, but reliable, digital presence/absence signals. The video sensor and the PIR sensor are built to cover the same field of view, in this way the PIR sensor can be aware of the the movements in the scene triggering the detection algorithm. The MSL is directly fitted into a PU board which is employed for digital image processing using single-cycle DSP instructions with configurable and flexible power management control. For example the typical current consumption for this microcontroller is about $1,7\,mA/MHz$ in RUN mode and only a few mA in SLEEP mode which is an attracting feature for wireless sensor networks design where the power consumption is a major constraint. Finally wireless communication is guaranteed by a Bluetooth transceiver adopted because of the bandwidth and the easy interface to host devices (i.e. PC, PDA). However, ZigBee radio interface is also supported.

The main goal of our system is to perform automatic detection of events such as the presence of abandoned and/or removed objects in the scene using non unobtrusive embedded platforms. Other specifications concern the need for low power consumption, the use of a PIR sensor to reduce the presence of false positives, and the possibility of sending an alarm to a remote host wirelessly. To satisfy the requirements, the information coming from the PIR sensor is used to ”wake up” the system in occurrence of specific events, as well as to evaluate when to start the video analysis stage. In fact, if the PIR sensor does not identify any event, the camera is switched off and the microcontroller is set to SLEEP mode minimizing the power consumption.

Fig. 3 shows the flow chart of the application. When triggered by an event from the PIR sensor, the system switches to RUN mode the ARM core, which runs full speed and all clocks are on, while the camera is kept off until movements in the field of view disappear. Then the camera is activated and takes a picture of the environment which is processed by the detection application, described in Section III, then the system switches back into SLEEP mode where the power consumption decreases up to 90% since only the PIR module operates as reported in next sections. This way the number of false positives is minimized beacuse the system processes the frames only in absence of moving objects in the monitored area enhancing robustness and autonomy. Finally, when an object is recognized as abandoned or removed, the system sends wirelessly alarms containing the number of objects, the regions of interest, size... and the full picture if requested by the host. In power characterization presented in this work, we considered a Bluetooth interface and we decided to send the full content of the image in order to estimate the autonomy of the platform.

III. The video analysis algorithm

This section describes the video analysis algorithm which is applied every time the intrusion detection block based on the PIR sensor detects absence of movements in the monitored scene and captures a new image from the scene activating the camera. By means of the PIR sensor, we can assume that all visible changes appearing in the scene in absence of movements have to be considered possible instances of removed or abandoned objects. Hence, a first stage of the algorithm consists in a background subtraction approach aimed at detecting visible changes in the scene background. Then, a labeling algorithm is implemented to enumerate and locate the

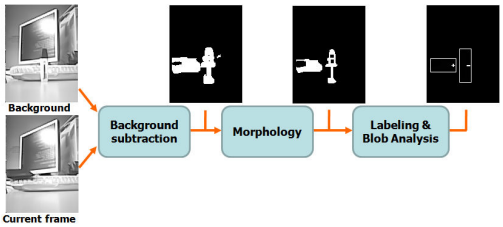

Fig. 4. The flow diagram of the proposed change detection algorithm.

areas of the image, or Regions-of-Interest (ROIs), where a stationary change of the background has taken place. Finally, a blob analysis stage provides the classification of each ROI between abandoned and removed object. All stages of the proposed video analysis algorithm have to be particularly memory efficient and need to avoid the use of floating point instructions given their implementation on the embedded architecture. Figure 4 shows the flow diagram of the algorithm.

a) Background subtraction: To detect stationary visible changes in the scene, we adopt a typical *background subtraction* approach, that is we compare the current frame captured from the camera, F, with a model of the background of the scene, B, computed at initialization time. To do this, each pixel at coordinates (x, y) in the current frame is compared with its homologous in the background model by means of a function aimed at measuring the similarity between the two image points.

To deal with illumination changes and photometric distortions that typically occur in real working conditions and may easily be misinterpreted as structural changes, we compute the Normalized Cross-Correlation (NCC) [23], which is invariant to linear photometric transformations between corresponding windows on F and B, on a squared neighborhood (i.e. a window of radius r) centered on the pixel under evaluation:

$$NCC(x,y) = \frac{F(x,y) \circ B(x,y)}{\| F(x,y) \|_2 \cdot \| B(x,y) \|_2} \qquad (1)$$

where the term at numerator is the dot product between $B(x, y)$ and $F(x, y)$, and the two terms at denominator represent the L_2 norms of $F(x, y)$ and $B(x, y)$, respectively.

Then, the NCC function is thresholded yielding a binary image, referred to as *change mask*, C, which highlights those parts of the current frame which have been subject to a change with respect to the background model:

$$C(x,y) = \begin{cases} changed, & NCC(x,y) < \tau_{NCC} \\ unchanged, & otherwise \end{cases} \qquad (2)$$

The use of the NCC is motivated by the fact that the system ought to be robust toward these kinds of distortions which can typically be found since the background model is computed once at initialization. On the other side, the implementation of the NCC function is particularly simple compared to more advanced approaches, and this aspect is particularly relevant since the algorithm has to be implemented on an ARM-based embedded architecture using a fixed point approach to maximize performance. In particular, to perform the square root and division operations of (1) a fixed-point square root function for ARM and a integer division have been utilized.

A typical effect of the use of the NCC over a window is that the segmentation of the foreground in the change mask becomes less accurate along the borders of the objects. In particular, there's a typical *fattening* effect, that is the object appears bigger since its borders are increased by a number of pixels proportional to r. To deal with this effect, a simple binary morphology operator of erosion is applied on the change mask as many times as the chosen value of r.

b) Labeling: After the background subtraction stage, a labeling algorithm is applied to group together connected components of the change mask. In this case, we use the algorithm proposed in [24], which is an efficient algorithm with low memory requirements for the labeling of binary images. In particular, the algorithm only requires two image scans and it has a memory complexity of $O(1)$. Once the labeling is performed, another image scan is deployed to compute the ROI coordinate of each connected component. Then a simple area-closing approach is performed to eliminate spurious components that might have been generated by noise.

c) Blob analysis: In the last stage of the algorithm, each valid ROI is classified either as an abandoned or removed object. The key idea beyond the adopted classification algorithm is that if an object is abandoned on the background, in F the number of edges along the borders of the corresponding connected component should increase compared to B. Conversely, if an object is removed from the background model, then F should display much less edges along the borders of the area where the object was initially located compared to B.

Hence, the approach relies on the estimation of the number of edges that appear on F along the borders of the connected component we want to classify. First of all, we detect all *contour* points within the ROI as those points that belong to the foreground and have at least one of their 8-connected neighbors set as background. On each contour point of coordinate (x, y), we compute the horizontal and vertical derivatives D_x, D_y of point $F(x, y)$ by means of the Sobel operator Then, we approximate the magnitude of the gradient in (x, y) as:

$$|G(x,y)| = max\left(|D_x(x,y)|, |D_y(x,y)|\right) \qquad (3)$$

A threshold is used to classify the contour point as being or not in presence of an edge in F. Then, the number of contour points associated with edges, N_{CE} is computed and thresholded:

$$Class(x,y) = \begin{cases} removed, & N_{CE} < \tau_C \\ abandoned, & otherwise \end{cases} \quad (4)$$

to yield final classification of the ROI.

IV. Experimental results

The above-mentioned application was fully implemented in ARM9 firmware. In the following we will focus on video sensor node power and performance. Since for this work we used only the internal 96KB SRAM, the camera is set to grab a 160x120 pixel (QCIF) gray scale image in YCbCr 4:0:0 format. The amount of byte for one image in this format is only 19200bytes, since each pixel uses only a byte. The abandoned/remove algorithm needs at least 3 images to work properly. In fact we need a stored background to achieve the NCC background subtraction and two images to store the change mask and the eroded image. For this reason, the total amount of RAM to stored all the required images grows up 76800bytes.

Power consumption is reported in Table Ia) while Table Ib) depicts also the processing time necessary to discriminate if objects are abandoned or removed from the environment. The time to elaborate the blob analysis depends on the number and size of ROIs. So it will be zero if the system does not detect any blob and about 100 ms for three ROIs 16x16. These results show how the power consumed by the whole system in SLEEP mode is less than 10% of power requirements of a fully active node. So without the information of a low-cost PIR sensors, the systems would waste the 90% of its energy, in the worst case. Moreover through PIR sensor information, the platform is able to switch on the camera as late as possible, reducing the camera power consumption again of around 20%. Moreover, the power consumption of wireless communication is minimized because of higher accuracy of the detection reduces the number of false positive.

To perform a quantitative evaluation of the abandoned/removed object detection algorithm, a dataset of images was acquired under real conditions within two sessions which differ by location and illumination conditions. A total of 50 images has been collected, each one showing different objects and simulating the frame collected by the system when the camera is switched on. In particular, each image includes a number of abandoned/removed objects that varies between one and three. Different tests with different backgrounds, chosen among the images of the dataset, have been performed, for a total of 141 cases of abandoned/removed objects tested (70 abandoned objects, 71 removed objects). Figure 5 shows a subset of the dataset.

(a) Power consumption of the video sensor node.

Component	Power $[mW]$
ARM9 mode (RUN / IDLE / SLEEP)	450 / 49,5 / 15
Video sensor mode (ON / IDLE)	165 / 23
TX/RX mode (ACTIVE / IDLE)	98 / 10
PIR sensor	1,5
Video Node	
Active with/without video sensor	626,5 / 484,5
Alarms Transmission	572,5
SLEEP, only PIR is Active	51

(b) Energy requirement of each task.

Task	Energy $[mJ]$	Time $[ms]$
Frame Acquisition	58,5	93,5
NCC Background Subtraction	455,8	940
Labeling	29	60
Blob Analysis	0 - 48,6	0 - 95
Image 160x120 Transmission	601,1	1050

TABLE I. Energy requirements of the low-power video system.

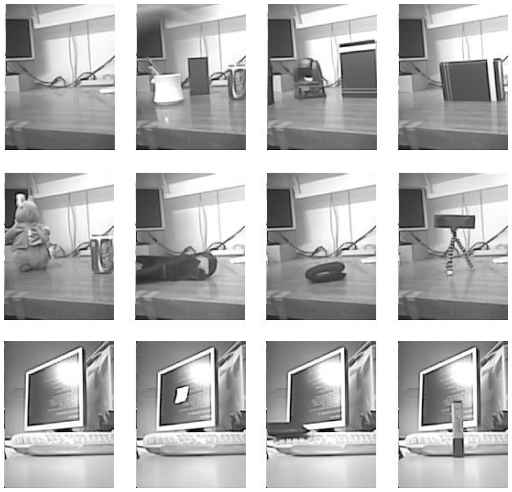

Fig. 5. Subset of the dataset used for the experimental evaluation.

In terms of change detection, our algorithm detected a total of 162 objects. In particular, it was always able to detect the presence of objects placed in the scene, with a percentage of false negatives (missed detections) equal to 0%. Instead, there's a number of false positives (false alarms) equal to 13% of the total number of detected objects.

To evaluate the fixed point approach we used the same datasets of images to compare the changed mask obtained from a NCC implementation on a floating-point Pentium4 architecture and on the presented fixed-point ARM-based solution. The difference concerns only 1% of the number of the pixels pointed out from the NCC implementation on

a PC. However, after the morphology operator of erosion, the accuracy of ROI detection on fixed-point ARM is not degraded with respect to the implementation on a Pentium4.

As for the performance reported by the classification algorithm, it yielded a number of misclassified objects equal to 7.8%. In particular, the percentage of correct detection for the removed object class is 98.6%, while the percentage of correct detection for the abandoned object class is 85.7%.

V. Conclusions

The interest in low-cost and small size video surveillance systems able to collaborate in networks of detection systems has been increasing over the last years. In this paper we have presented a multi-modal video sensor node designed for low-power and low-cost video surveillance which is able to detect objects abandoned or removed in the environment. The system is multi-modal and a PIR sensor assists a CMOS video camera to increase the efficiency of the algorithm and to extend the life time of the system. We addressed different configurations and characterized the system in terms of runtime execution, power consumption and efficiency.

Acknowledgements

The work presented in this paper has been founded by the European Network of Excellence ArtistDesign and by SCALOPES, an Artemisia JU project. In addition, the authors would like to thank STMicroelectronics for the hardware support.

References

[1] S. Lu, J. Zhang, and D. Feng, "A knowledge-based approach for detecting unattended packages in surveillance video," in *Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS 06)*, 2006.

[2] S. Lim and L. Davis, "A one-threshold algorithm for detecting abandoned packages under severe occlusions using a single camera," CS Dept., University of Maryland, Tech. Rep. CS-TR-4784, 2006.

[3] S. Ferrando, G. Gera, and C. Regazzoni, "Classification of unattended and stolen objects in video-surveillance system," in *Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS 06)*, 2006.

[4] J. San Miguel and J. Martnez, "Robust unattended and stolen object detection by fusing simple algorithms," in *Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS 08)*, 2008, pp. 18–25.

[5] M. Bhargava, C. Chen, M. Ryoo, and J. Aggarwal, "Detection of abandoned objects in crowded environments," in *Proc. of IEEE Conf. on Advanced Video and Signal Based Surveillance (AVSS07)*, 2007, pp. 271–276.

[6] F. Porikli, Y. Ivanov, and T. Haga, "Robust abandoned object detection using dual foregrounds," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, 2008.

[7] M. Spengler and B. Schiele, "Automatic Detection and Tracking of Abandoned Objects," in *Proc. IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2003.

[8] Y. e. a. Tian, "Robust and efficient foreground analysis for real-time video surveillance," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 05)*, 2005, pp. 1182–1187.

[9] P. Spagnolo, A. Caroppo, M. Leo, T. Martifiggiano, and T. D'orazio, "An abandoned/removed objects detection algorithm and its evaluation on pets datasets," in *Proc. IEEE Conf. on Advanced Video and Signal Based Surveillance (AVSS 06)*, 2006, p. 17.

[10] M. Beynon, D. Van Hook, M. Seibert, A. Peacock, and D. Dudgeon, "Detecting abandoned packages in a multi-camera video surveillance system," in *Proc. IEEE Conf. on Advanced Video and Signal Based Surveillance (AVSS 03)*, 2003, pp. 221–228.

[11] C. Sacchi and C. Regazzoni, "A distributed surveillance system for detection of abandoned objects in unmanned railway environments," *IEEE Trans. Vehicular Technology*, vol. 49, no. 5, 2000.

[12] N. Bird, S. Atev, N. Caramelli, R. Martin, O. Masoud, and N. Papanikolopoulos, "Real time, online detection of abandoned objects in public areas," in *Proc. IEEE Conf. on Robotics and Automation (ICRA06)*, 2006, pp. 3775 – 3780.

[13] M. Magno, F. Tombari, D. Brunelli, L. Di Stefano, and L. Benini, "Multi-modal video surveillance aided by pyroelectric infrared sensors," in *Proc. ECCV Workshop on Multi-camera and Multi-modal Sensor Fusion, Algorithms and Applications (M2SFA2)*, 2008.

[14] P. de la Hamette et al., "Architecture and applications of the fingermouse: a smart stereo camera for wearable computing hci," *Personal Ubiquitous Comput.*, vol. 12, no. 2, pp. 97–110, 2008.

[15] D. Corley and E. Jovanov, "A low power intelligent video-processing sensor," in *Proc. Thirty-Fourth Southeastern Symposium on System Theory*, 2002, pp. 176–178.

[16] D. Li, Y. Jiang, and G. Chen, "A low cost embedded color vision system based on sx52," in *Proc. IEEE International Conference on Information Acquisition*, 2006, pp. 883–887.

[17] G. Zhang, T. Yang, S. Gregori, J. Liu, and F. Maloberti, "Ultra-low power motion-triggered image sensor for distributed wireless sensor network," in *Proc. IEEE Sensors*, vol. 2, 2003, pp. 1141–1146 Vol.2.

[18] W.-c. Feng, B. Code, E. Kaiser, M. Shea, W.-c. Feng, and L. Bavoil, "Panoptes: scalable low-power video sensor networking technologies," in *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*. New York, NY, USA: ACM, 2003, pp. 562–571.

[19] L. Ferrigno and A. Pietrosanto, "A low cost visual sensor node for bluetooth based measurement networks," in *Proc. 21st IEEE Instrumentation and Measurement Technology Conference IMTC 04*, vol. 2, 2004, pp. 895–900 Vol.2.

[20] P. Garda, O. Romain, B. Granado, A. Pinna, D. Faura, and K. Hachicha, "Architecture of an intelligent beacon for wireless sensor networks," in *Proc. NNSP'03 Neural Networks for Signal Processing 2003 IEEE 13th Workshop on*, 2003, pp. 151–158.

[21] T. e. a. He, "Vigilnet: An integrated sensor network system for energy-efficient surveillance," *ACM Trans. Sen. Netw.*, vol. 2, no. 1, pp. 1–38, February 2006.

[22] A. Rajgarhia, F. Stann, and J. Heidemann, "Privacy-sensitive monitoring with a mix of IR sensors and cameras," in *Proceedings of the Second International Workshop on Sensor and Actor Network Protocols and Applications*, August 2004, pp. 21–29.

[23] F. Tombari, L. Di Stefano, and S. Mattoccia, "A robust measure for visual correspondence," in *Proc. Int. Conf. on Image Analysis and Processing*, 2007, pp. 376–381.

[24] L. Di Stefano and A. Bulgarelli, "A simple and efficient connected components labeling algorithm," in *Proc. Int. Conf. on Image Analysis and Processing (ICIAP 99)*, 1999, pp. 322–327.