Basic Video-Surveillance with Low Computational and Power Requirements Using Long-Exposure Frames

Vincenzo Caglioti and Alessandro Giusti

Dipartimento di Elettronica e Informazione, Politecnico di Milano
P.za Leonardo da Vinci, 32 20133 Milano – Italy
alessandro.giusti@polimi.it, vincenzo.caglioti@polimi.it

Abstract. Research in video surveillance is nowadays mainly directed towards improving reliability and gaining deeper levels of scene understanding. On the contrary, we take a different route and investigate a novel, unusual approach to a very simple surveillance task – activity detection – in scenarios where computational and energy resources are extremely limited, such as Camera Sensor Networks.

Our proposal is based on shooting long-exposure frames, each covering a long period of time, thus enabling the use of frame rates even one order of magnitude slower than usual – which reduces computational costs by a comparable factor; however, as exposure time is increased, moving objects appear more and more transparent, and eventually become invisible in longer exposures. We investigate the consequent tradeoff, related algorithms and their experimental results with actual long-exposure images. Finally we discuss advantages (such as its intrinsic ability to deal with low-light conditions) and disadvantages of this approach.

1 Introduction

In this paper we introduce a novel low-level technique for implementing basic video surveillance functionality, whose main advantage is the minimal amount of computational effort required; this is achieved through the use of long-exposure frames and their analysis through basic image processing techniques.

The main application scenario is Wireless Sensor Networks and comparable systems, which are recently receiving much attention as an innovative computing platform enabling novel, powerful applications, many of which are related to video surveillance. In wireless sensor networks, individual sensors are designed to be as simple, tiny and cheap as possible and have minimal computing and communication capabilities [3]; moreover, power consumption is usually a critical factor, as each sensor is powered by a battery which must last as long as possible (as replacement is usually unpractical, and often not even foreseen); therefore, attention to avoiding unnecessary computational operations is compulsory, and directly affects a critical parameter such as the lifetime of the system.

Cameras are usually problematic in wireless sensor networks as they generate huge amounts of data when compared to temperature, light, humidity, and magnetism sensors, which are currently mainstream in the field. On the other hand,

J. Blanc-Talon et al. (Eds.): ACIVS 2008, LNCS 5259, pp. 218–229, 2008.

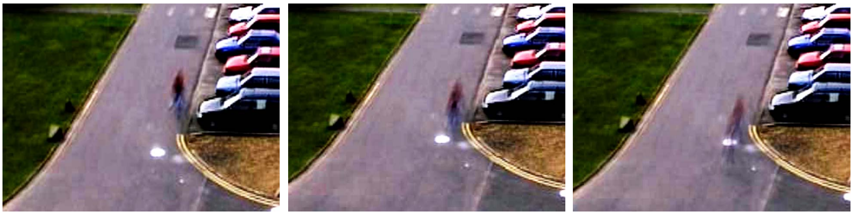


Fig. 1. Simulated long exposure images in a typical surveillance environment (crops of larger images, contrast-stretched). As the exposure time increases (from left to right: 1, 2 and 3 seconds), the moving subject appears more and more blurred and transparent.

Fig. 2. Nodes of wireless sensor networks and wireless camera networks, battery powered. Only a very careful use of their limited computational and communication resources can allow them to reach a years-long lifetime.

cameras have several important advantages w.r.t. simpler sensors especially in surveillance and scene understanding applications, as they yield a much higher informational content. Moreover, as basic image sensors become cheaper, the total cost of an embedded camera can be very tolerable nowadays[1].

Therefore, a fair amount of research is recently being devoted to camera-equipped wireless sensor networks, named *camera sensor networks*[2]; these systems are employed not only for surveillance [13] but also other applications, such as structural health monitoring of buildings [2], human behavior recognition [12], people and object tracking [6,5]. One shared challenge in these systems is the need to reduce the amount of data to be transmitted, which implies local processing/storage of images [7,10]. Several papers have also challenged the problem of calibration of cameras in camera sensor networks [11], in order to recover their relative position.

[1] For surveillance applications, cost and size may be further reduced by exploiting pinhole cameras which do not even need optics: one disadvantage of pinhole cameras is that they allow minimal amounts of light to reach the sensor; this is not a problem in our setting as we deal with long-exposure images. Moreover, in our scenario electronics do not need to support fast acquisition rates.

[2] Not to be confused with *wireless camera networks*, which usually indicate networks of bigger and more powerful cameras whose main goal is to wirelessly transmit their data to a central collection point.

In this paper, we describe our low-level technique in the context of a typical camera sensor network, which is introduced in the next section; we then describe the main motivations of our approach. In Section 3 we recall the image formation model for long-exposure frames, then show how activity can be detected from them (Section 4). Section 5 presents implementation notes and Section 6 summarizes our experimental results. The main advantages, disadvantages and possible improvements of our technique are finally discussed in Section 7.

2 A Straightforward Surveillance System Based on a Camera Sensor Network

In the following, we will refer to a straightforward surveillance system based on a simple camera sensor network; the network has two operation modes:

- an “idle” mode, which is expected to dominate the great majority of the network’s lifetime, in which each sensor is only required to detect any ongoing activity with the least possible expense of resources;
- when any activity is detected, a message is propagated through the network which switches it to an “alert” mode, where more resources and sophisticated processing and decision making can be employed in order to assess the occurring event with better precision, possibly saving or transmitting images of the scene for documentation purposes or further processing.

In this paper we focus of the first mode, and provide a technique for determining activity with minimal consumption of resources.

We point out that in this phase, processing the images on the sensor itself is a strict requirement. In fact, regularly transmitting acquired frames to a more powerful computational entity (base station) is prohibitive due to the power consumption implied by the amount of data to send; in a realistic setting, a sensor may not be able to transmit more than several hundreds of frames before draining its batteries. Even when considering compression, which has a significant computational cost in this platform, it is unrealistic to assume a sensor operating continuously for more than several days or weeks. As this may be viable for the “alert” mode, which is rarely triggered for short amounts of time, it is not for the phase we are considering.

2.1 Frame Differencing for Activity Detection

A naive approach to detect activity during the system’s “idle” mode, apparently rather simple from the computational point of view, implies acquiring frames continuously and using frame differencing to detect movement [8]. Frame differencing itself requires a number of operations for each frame which is linear with the number of pixels; therefore, two factors directly influence computational complexity:

- camera resolution;
- frame rate.

The camera resolution, especially when used for a simple activity detection phase, is rarely critical and can be often reduced significantly; some sensors can be configured to return a downsampled image (binning); or, alternatively, the software can process only a subset of the available pixels.

Frame rates routinely for this task used range from 20 down to about 1 frames per second. Faster frame rates are obviously useless (or even counter-productive) in this setting; slower frame rates, on the other hand, leave holes in the temporal coverage of the scene which can negatively affect detection rate, if they are longer than the expected duration of the visible activity.

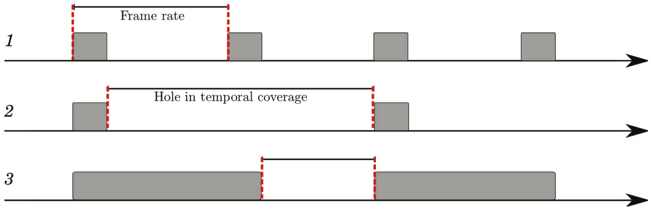


Fig. 3. A simplified representation of our approach: gray boxes represent the camera exposure time; lines 1 and 2 show how decreasing the frame rate causes larger holes in temporal coverage, which may prevent event detection. Line 3 shows a frame rate as low as line 2, but with a longer exposure time, which decreases the size of coverage holes.

We propose to shoot long-exposure frames to compensate the loss in temporal coverage due to a slower framerate, which allows us to reach very slow frame rates without compromising the detection probability. Unfortunately, moving objects (such as cars or people) in long-exposure images degrade to semitransparent motion smears, which makes them practically invisible in exposures of excessive length[4]. In the following section, we recall an image formation model which explains this phenomenon.

3 A Model for Motion Smears

A motion blurred image is obtained when the scene projection on the image plane changes during the camera exposure period $[t', t'']$. The resulting image C is the integration of an infinite number of sharp images, each exposed for an infinitesimal portion of $[t', t'']$. In an equivalent interpretation (Figure 4), we can consider the motion blurred image as the temporal average of infinite sharp images I_t, each taken with the same exposure time $t'' - t'$ but representing the scene frozen at a different instant $t \in [t', t'']$. This technique is implemented in many 3D rendering packages for accurate (but computationally expensive) synthesis of motion blurred images; we also exploit this property in our experiments (see Section 6 for simulating long-exposure images with public video datasets, such as [1].

If the camera is static and a single object is moving in the scene, the static background in the final image is sharp since its pixels are of constant intensity

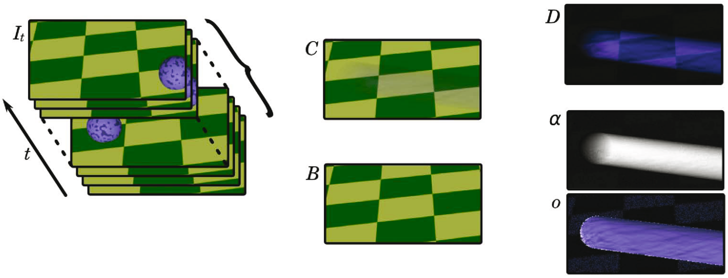


Fig. 4. Original image (C), background (B), transparency (α, rescaled for display purposes), foreground map (o). The image C of the motion blurred object can be interpreted as the temporal average over the exposure time of an infinite number of still images I_t (top). The blurred smear can be interpreted as a semitransparent layer, whose alpha matte (transparency) we analyze in the following.

in each I_t; conversely, the image pixels which are affected by the moving object, possibly only in some of the I_t images, belong to the motion-blurred image (smear) of the object.

For a pixel p, define $i(p) \subseteq [t', t'']$ the set of time instants during which p belongs to the object image. We finally define $\alpha(p)$ as the fraction of $[t', t'']$ during which the object projects to p:

$$\alpha(p) = ||i(p)||/(t'' - t'). \qquad (1)$$

Let $B(p)$ be the intensity of the background at p. Since C is the temporal average of all the I_t images, $C(p) = \alpha(p)o(p) + (1 - \alpha(p))B(p)$. $o(p)$ is the temporal average over $i(p)$ of the intensity of image pixel $C(p)$:

$$o(p) = \frac{1}{||i(p)||} \int_{t \in i(p)} I_t(p)\, dt. \qquad (2)$$

To sum up, the intensity of a pixel p in the motion blurred image C can be interpreted as the convex linear combination of two factors: the "object" intensity $o(p)$, weighted $\alpha(p)$, and the background intensity. The resulting equation is the well-known Porter-Duff alpha compositing equation [9] for a pixel with transparency $\alpha(p)$ and intensity $o(p)$ over the background pixel $B(p)$.

The object intensity $o(p)$ can be interpreted as the intensity that p would have in the motion blurred image over a black background, rescaled by a $\frac{1}{\alpha(p)}$ factor. $o(p)$ is meaningless if p is not affected by the object image during the exposure.

4 Detecting Moving Objects with Long-Exposure Frame Differencing

The model introduced in the previous section clearly shows in (1) that in general the opacity (alpha) values associated to the image of a moving object decrease as the exposure time is increased; this causes the object to become invisible to the naked eye in longer exposures. This phenomenon is sometimes exploited by experienced photographers for making rain invisible by increasing exposure time; in our situation, however, it reduces the visibility of moving objects, which may remain undetected.

As an example, we can consider a typical surveillance scenario and compute the theoretical transparency of the image of a person walking with constant speed in a direction parallel to the image plane. Depending on the setting and with some exceptions we will highlight later, the image of the person projects to a pixel for approximately 1/5 of a second. The alpha value of the person's image will be, in average, about $\alpha = \frac{0.2}{[t''-t']}$; in this case, if the exposure time is 4 seconds, the person's image will have a very limited opacity of $\alpha = 1/20$; whether it is actually visible depends on the contrast between the person's (sharp) image and the background. Considering an 8 bit image in the $[0\,255]$ range, and the maximum contrast between object and background, the person's smear will differ from the background by only 10-15 intensity values[3].

On the other hand, the image of the person extends to a bigger area than the area it would cover in a short-exposure image.

In general, ordinary frame differencing works rather well with long-exposure images for identifying changes between one frame and the following, when setting low detection thresholds.

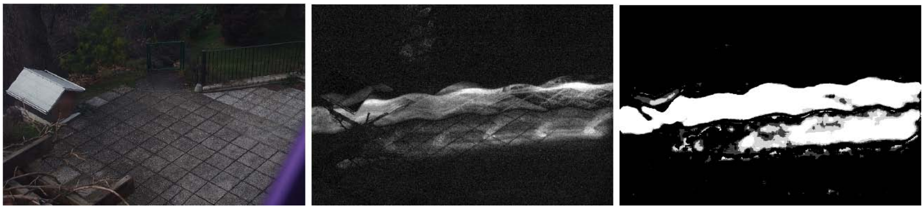


Fig. 5. A person running with a white jacket is almost invisible to the naked eye in a 4 seconds exposure (left); however, it is easily detected by naive frame differencing. Center: difference to previous frame. Right: after thresholding. Note the visible pattern of the running oscillations.

Some problems with this approach may arise due to slow global luminance changes in the whole scene; such events are very frequent especially in outdoor environments or scenes lit through windows, due to the low detection threshold and the significant inter-frame time; then even the natural motion of the sun may cause false triggers; a straightforward solution is computing the median ratio between corresponding pixels in the two images, and correcting the second

[3] Note that these considerations would not apply if the object would be over-exposed if still: then, its actual intensity would be higher than the sensor dynamic range; this is the reason why objects very bright with respect to their surroundings, such as stars or light sources at night, can leave highly visible streaks in exposures even some hours long.

image by means of this ratio. It can be efficiently implemented by computing the correction factor on a subset of the image pixels uniformly sampled over the whole image.

When using longer exposures requiring very low thresholds for detection, the reliability of the system can be improved by applying some processing to the computed deviations before the threshold is applied; efficiency constraints may instead suggest to apply it after thresholding on the obtained binary image; our experiments did not show any of these possibilities to be significantly better.

Finally, the proposed process looks as follows:

```
for each new frame $C_i$ do
    compute correction factor $h \leftarrow \text{median}(C_{i-1}(p)/C_i(p))$ with $p \subseteq$ all pixels
    compute differences image as $D \leftarrow h \cdot C_i - C_{i-1}$
    apply spatial median filtering to $D$
    threshold $D$ to $T$
    if $T$ has more than a preset number of white pixels then
        trigger alert state
    end if
end for
```

In our experiments, the threshold is global to the whole image, static, and automatically derived from the measured noise. However, many sophisticated approaches to foreground extraction are proposed in literature, which may further improve our results; they provide techniques for building and maintaining accurate background models, and setting adaptive thresholds for different parts of the image. We are currently evaluating whether they provide a sensible performance improvement in this context, and if the increased computational effort they require is justified.

5 Implementation Notes

In many cases, taking long exposures in daylight environments with ordinary customer equipment is not immediate as overexposure is sometimes unavoidable, because such sensors are not designed for long exposures in well-lit environments. Using an aperture as small as possible and if needed ND (Neutral Density) filters for reducing the amount of light reaching the sensor is viable solution to this problem.

Due to the very small sensor amplification required, sensor noise (ISO noise) is dramatically reduced, and allows meaningful comparisons involving even the least significant bits of every pixel's intensity value (assuming 8 bits per pixel). Therefore, choosing a threshold of 2 or 3 intensity values out if 255 works well even with small cheap sensors. On the other hand, long-exposure noise does not affect frame differencing as it creates artifacts not changing from frame to frame. In some cases, using a 12-bit dynamic range can help in identifying fainter smears for exposures longer than 30 seconds.

6 Experimental Results

In this section we provide experimental results on:

- actual long-exposure images taken with a number of different consumer digital cameras (figures 7, 10 and 8);
- simulated long-exposure images synthesized from publicly-available datasets such as [1] (figures 6 and 9). As anticipated in Section 3, the synthesized images are created by averaging many frames of an ordinary video shot at 20 frames per seconds.

They are a representative part of the whole set of experiments, which is composed by exposures ranging from 1 to 60 seconds, shot in 7 different scenarios, involving both people and cars.

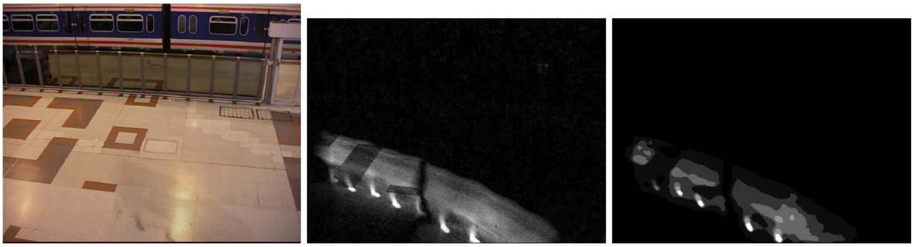


Fig. 6. A person walking in a public building (synthesized from video): 4 seconds exposure. Note that feet are highly visible in the difference image as they are fixed for longer periods of time. Also note that patterns in the background influence the difference image as they affect contrast to foreground.

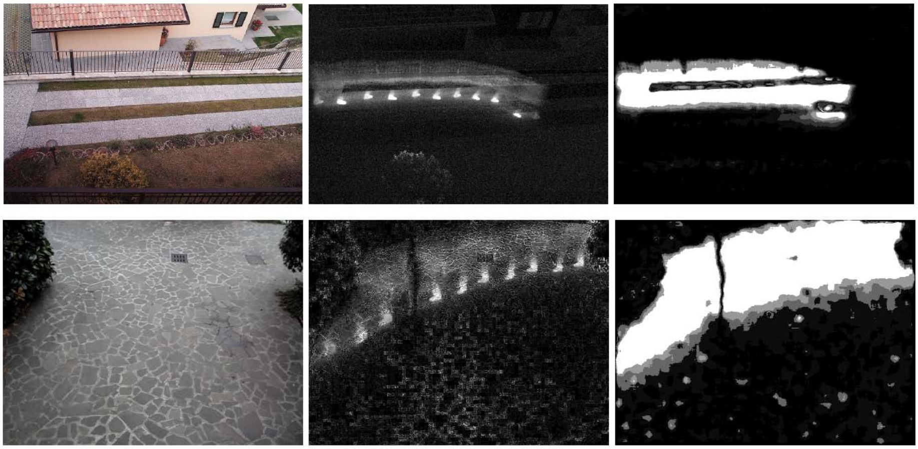


Fig. 7. A person walking in an outdoor setting. First row: 4 seconds exposure time; Second row: 6 seconds.

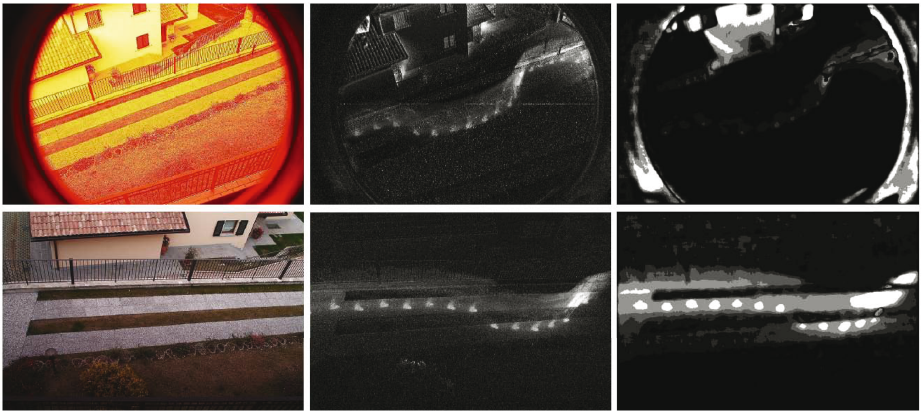


Fig. 8. 20 seconds exposure

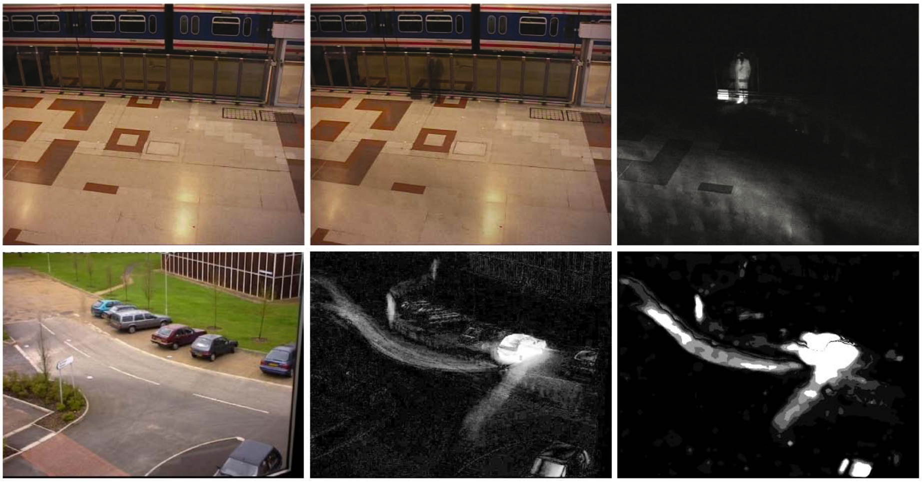
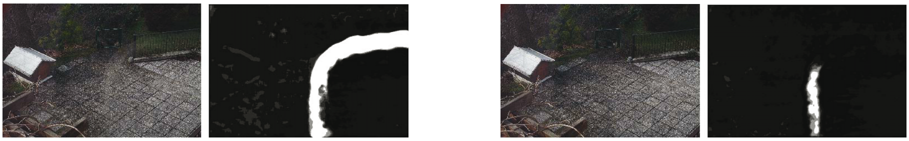


Fig. 9. *Up*: 15 seconds exposure; note path of person and its clear silhouette where he stood still; his image is also visible, semitransparent, in the long-exposure frame. *Down*:5 seconds exposure; note path of car while parking, and trace of nearby motion. Images syntesized from public videos[1].

In all tested scenarios, exposures up to 6 seconds allowed safe and very reliable detection of any activity without false positives, when using thresholds automatically derived from the measured image variance. In most scenarios, however, exposure times up to 15-25 seconds can be reached while keeping very reliable operation, provided that the background has a sufficient contrast with the subject.



(a) subject with a white jacket (higher contrast with the background)

(b) subject with a black jacket (lower contrast with the background)

Fig. 10. The detection is harder when the subject has a low contrast with the background, and when the subject moves in a direction parallel to the image plane, as its image overlaps with a given pixel for a shorter time

7 Discussion, Future Works and Conclusions

We demonstrated that using long-exposure frames allows a surveillance system to greatly reduce its frame rate without compromising surveillance ability; this is a very important achievement for e.g. wireless sensor network systems, which are based on nodes with minimal computing power and limited energy resources. Although blurred smears of moving objects are often invisible to the naked eye even for exposures shorter than 2-3 seconds, ordinary frame differencing makes practical detection of a walking person possible in exposures up to 15-25 seconds, depending on the scenario; if implemented correctly, this simple approach may lead to power and computational requirements even one order of magnitude lower than with traditional methods.

The approach is technically practical because the long exposure requires very low sensor amplification, which translates to very clean and noise-free images: this allows low detection thresholds to be set; spatial filtering can also be employed to improve detection of faint smears, at the expense of an increased computational effort, which may often be very problematic in the considered scenario.

Other than our main motivation, i.e. greatly reduced computational and power requirements, our technique also has some other advantages over using ordinary short-exposure frames:

- it naturally works in low light, which is a very significant advantage in surveillance applications;
- it is rather robust to false-alarms triggered in ordinary systems by rapidly changing pixels, due to periodically moving objects as fans and trees in the wind, or their shadows; in fact, these intensity variations are inherently averaged over time in a long-exposure image, and often cause negligible differences between frames. Although we noted some hints of this interesting property in our experiments, we still have to determine whether the practical implications can be beneficial in practice.

On the other hand, an ineliminable disadvantage of our approach is its reaction time, as the detection is deferred to the end of the exposure.

Activity detection, which is the simple task we deal with in this work, may also be easily solved using different, simpler sensors. However, our approach has the advantage of exploiting the same camera which can be used for more sophisticated purposes in the "alert" mode of the system.

Moreover, as the feasibility and utility of the approach is now demonstrated, we now plan to extract some additional information contained in the blurred image which should be accessible by means of relatively simple image processing techniques.

- The path of the object is a very summarized information which may be easily transmitted to other sensors for information integration; it can be easily extracted as the centerline of the smear.
- as shown in figures 6 and 7, walking people leave visible gait patterns, which may be exploited to discern people from other moving objects.

Acknowledgments

Students Marco Uberti and Andrea Riva contributed to the experimental part of this work.

References

1. Pets, performance evaluation of tracking and surveillance, `http://www.cvg.rdg.ac.uk/slides/pets.html`
2. Basharat, A., Catbas, N., Shah, M.: A framework for intelligent sensor network with video camera for structural health monitoring of bridges. In: PERCOMW 2005: Proceedings of the Third IEEE International Conference on Pervasive Computing and Communications Workshops, Washington, DC, USA, pp. 385–389. IEEE Computer Society, Los Alamitos (2005)
3. Crossbow. Mica2 data sheet
4. Giusti, A., Caglioti, V.: Isolating motion and color in a motion blurred image. In: Proc. of British Machine Vision Conference (BMVC) 2007 (2007)
5. Hengstler, S., Aghajan, H.: Application-driven design of smart camera networks. In: Proceedings of the COGnitive systems with Interactive Sensors (2007)
6. Kulkarni, P., Ganesan, D., Shenoy, P., Lu, Q.: Senseye: a multi-tier camera sensor network. In: MULTIMEDIA 2005: Proceedings of the 13th annual ACM international conference on Multimedia, pp. 229–238. ACM Press, New York (2005)
7. Mathur, G., Chukiu, P., Desnoyers, P., Ganesan, D., Shenoy, P.: A storage-centric camera sensor network. In: SenSys 2006: Proceedings of the 4th international conference on Embedded networked sensor systems, pp. 337–338. ACM Press, New York (2006)
8. Migliore, D.A., Matteucci, M., Naccari, M.: A revaluation of frame difference in fast and robust motion detection. In: VSSN 2006: Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks, pp. 215–218. ACM, New York (2006)
9. Porter, T., Duff, T.: Compositing digital images. Computer Graphics (1984)
10. Rahimi, M., Baer, R., Iroezi, O.I., Garcia, J.C., Warrior, J., Estrin, D., Srivastava, M.: Cyclops: in situ image sensing and interpretation in wireless sensor networks. In: SenSys 2005: Proceedings of the 3rd international conference on Embedded networked sensor systems, pp. 192–204. ACM, New York (2005)

11. Rekletis, I.M., Dudek, G.: Automated calibration of a camera sensor network. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, Edmonton Alberta, Canada, August 2-6, pp. 401–406 (2005)
12. Teixeria, T., Lymberopoulos, D., Culurciello, E., Aloimonos, Y., Savvides, A.: A lightweight camera sensor network operating on symbolic information. In: SenSys: Proceedings of the Workshop on Distributed Smart Cameras (2006)
13. Valera, M., Velastin, S.: Intelligent distributed surveillance systems: a review. In: IEE Proceedings on Vision, Image and Signal Processing (2005)