Image Mining for Infomobility

Massimo Magrini[1], Davide Moroni[1], Christian Nastasi[2], Paolo Pagano[2]
Matteo Petracca[2], Gabriele Pieri[1], Claudio Salvadori[2] and Ovidio Salvetti[1]

[1] Institute of Information Science and Technologies (ISTI)
Italian National Research Council (CNR), Pisa, Italy

[2] ReTis Laboratory (CEIIC)
Scuola Superiore SantAnna, Pisa, Italy

Abstract. The wide availability of embedded sensor platforms and low-cost camera sensors – together with the developments in wireless communication – make it now possible the conception of pervasive intelligent systems based on *vision*. Such systems may be understood as distributed and collaborative sensor networks, able to produce, aggregate and process images in order to mine the observed scene and communicate the relevant information found about it. In this paper, we investigate the peculiarities of visual sensor networks with respect to standard vision systems and we identify possible strategies to tackle the image mining problem. We argue that multi-node processing methods may be envisaged to decompose a complex task into a hierarchy of computationally simpler problems to be solved over the nodes of the network. We illustrate these ideas by describing an application of visual sensor network to infomobility.

1 Introduction

In the last years, the wide availability of embedded systems and low-cost camera sensors – together with the developments in wireless communication – has made it possible to conceive sensor-based pervasive intelligent systems centered on image data [1]. Such visual Wireless Sensor Networks (WSNs), employing a great number of low power camera nodes, may support a large class of novel vision-based applications thanks to the great informative power of imaging. For example, visual WSNs may be used to monitor in real-time the crowd in shopping mall, airports and stadiums. By mining the scene, the network may detect anomalous – and, thus, potentially dangerous – events [2]. Similarly, visual WSNs may be used for environmental monitoring, for the remote control of elderly patients in e-Health [3] and for human gesture analysis in ambient intelligence applications [4].

In brief, the key feature of visual WSNs is the combination of the versatility and independence from a physical infrastructure typical of general WSNs with the richness of information that can be gained through imaging techniques, computer vision and image understanding. In this sense, a visual WSN may be understood as a distributed and collaborative sensor network, able to produce, aggregate and process images in order to mine the observed scene and communicate the relevant information found about

Magrini M., Moroni D., Nastasi C., Pagano P., Petracca M., Pieri G., Salvadori C. and Salvetti O. (2010).
Image Mining for Infomobility.
In *Proceedings of the Third International Workshop on Image Mining Theory and Applications*, pages 35-44
DOI: 10.5220/0002962000350044
Copyright © SciTePress

it. In particular, each acquisition node is capable to acquire a view of the scene to be mined. Then, in cooperation with each other, the various devices in the network are able to process and aggregate the acquired views in order to predicate something about the observed scene.

A successful design and development of such a system cannot be achieved without suitable solutions to the involved computer vision problems. Although the computer vision problems may be still decomposed into basic computational tasks (such us feature extraction, object detection and object recognition), in the context of visual WSNs, it is not directly possible to use all the methods that have been developed to solve such tasks and that are already available in the specific literature [5]. Indeed, since WSNs usually require a large number of sensors, possibly scattered over a large area, the unit cost of each device should be small to make the technology affordable. Therefore, cost constraints limit the computational and transmission power of sensor nodes as well as the fidelity of acquired images, with consequences on the employable computer vision algorithms. Since untethered sensors are usually supplied by batteries, it is also necessary to analyze carefully the tradeoffs between quality of processing and energy consumption, in order to avoid too frequent battery replacement.

In light of these considerations, performing computer vision tasks over WSNs is somewhat challenging from a technological viewpoint. However, probably, the change in perspective offered by the possibility to perform pervasive computing and by the ductility in organizing the network topology greatly compensates for the current technical limits of visual WSNs and is even more intriguing, opening the way to more theoretical investigations. In particular, in this paper, we investigate multi-node processing methods that may be envisaged to decompose a complex vision task into a hierarchy of computationally simpler problems, to be solved over the nodes of the network. Besides computational advantages, such hierarchical decision making approach may lead to more robust and fault-tolerant results. We illustrate these ideas by describing an application of visual sensor network to infomobility. To this end, we consider an experimental setting in which several views of a parking lot are acquired by the sensor nodes in the network. By integrating the various views, the network is capable to provide a description of the scene in terms of the available spaces in the parking lot.

2 Background

Embedded vision platform first appeared in connection with video-surveillance and robotics. Thanks to the drop in technology costs, nowadays, their application range has become wider so as to encompass several sectors of public and private life and, in particular, infomobility. We first discuss existing embedded vision platforms and basic features of visual WSNs (Section 2.1); then we survey the specificities of performing image analysis over WSNs, introducing in particular the multi-view vision problem (Section 2.2).

2.1 Visual Sensor Networks

Following the trends in low-power processing, wireless networking and distributed sensing, visual WSNs are experiencing a period of great interest, as shown by the recent

scientific production (see e.g [6]). A visual WSN consists of tiny visual sensor nodes called camera nodes, which integrate the image sensor, the embedded processor and a wireless RF transceiver [1]. The large number of camera nodes forms a distributed system where the camera nodes are able to process image data locally (*in-node processing*) and to extract relevant information, to collaborate with other-cameras – even autonomously – on the application specific task, and to provide the system user with information-rich descriptions of the captured scene.

In the last years, several research projects produced prototypes of embedded vision platforms which may be deployed to build a visual WSN. Among the first experiences, Panoptes project [7] aimed at developing a scalable architecture for video sensor networking applications. However, the size of the sensor, its power consumption, its relatively high computational power and storage capabilities makes Panoptes sensor more akin to smart high-level cameras than to untethered low-power low-fidelity sensors. In the MeshEye project [8], an energy-efficient smart camera mote architecture was designed, mainly with intelligent surveillance as target application. MeshEye mote has an interesting special vision system based on a stereo configuration of two low-resolution low-power cameras, coupled with a high resolution color camera. Another interesting example of low-cost embedded vision system is represented by the CMUcam3 [9], developed at the Carnegie Mellon University. More precisely, the CMUcam3 has been specially-designed to provide an open-source, flexible and easy development platform with robotics and surveillance as target applications. More recently, the CITRIC platform [10], once equipped with a standard RF transceiver, is suitable for the development of visual WSN. Indeed, the CITRIC system allows to perform moderate image processing task *in-network*, that is along the nodes of the network. In this way, there are less stringent issues regrading transmission bandwidth than with respect to centralized solutions.

2.2 Image Analysis Over Visual WSNs

Gathering information from a network of scattered cameras, possibly covering a large area, is a common feature of many videosurveillance and ambient intelligence systems. However, most of classical solutions are based on a centralized approach, in which image processing is accomplished in a single unit. In this respect, the need to introduce distributed intelligent system – where distribution is not merely physical but also semantic – is motivated by several requirements, namely [11]:

- *Speed:* in-network distributed processing is inherently parallel; in addition, the specialization of modules permits to reduce the computational burden in the high level decisional nodes.
- *Bandwidth:* in-node processing permits to reduce the quantity of transmitted data, by transferring only information-rich parameters about the observed scene and not the redundant image data stream.
- *Redundancy:* a distributed system may be re-configured in case of failure of some of it components, still keeping the overall functionalities.
- *Autonomy:* each of the nodes may process the images asynchronously and may react autonomously to the perceived changes in the scene.

In particular, these issues suggests to move a part of intelligence towards the camera nodes. In these nodes, artificial intelligence and computer vision algorithms are able to provide autonomy and adaptation to *internal conditions* (e.g. hardware and software failure) as well as to *external conditions* (e.g. changes in weather and lighting conditions).

The mining and deployment of visual information involve particular problems in computer vision, such as change detection in image sequences, object detection, object recognition, tracking, and image fusion for multi-view analysis. For each of this problems, there exists a large corpus of already implemented methods (see e.g. [12] for a survey of change detection algorithms); however most of the techniques currently available are not suitable to be used in visual WSNs, due to the high computational complexity of algorithms or to excessively demanding memory requirements. Nevertheless, some attempts to employ non-trivial image analysis methods over visual WSN have been done. For example, [13] presents a visual WSN able to support the query of a set of images in order to search for a specific object in the scene. To achieve this goal, the system uses a representation of the object given by the Scale Invariant Feature Transform (SIFT) descriptors [14]. SIFT descriptors are known to support robust identification of objects even among cluttered background and under partial occlusion situations, since the descriptors are invariant to scale, orientation, affine distortion and partially invariant to illumination changes. In particular, using SIFT descriptors allows retrieving the object of interest from the scene, no matter at which scale it is imaged.

Interesting computer algorithms are also provided on the CMUcam3 vision system [9]. Besides basic image processing filters (such as convolutions), methods for real-time tracking of blobs on the base either of color homogeneity or frame differencing are available. A customizable face detector is also included. Such detector is based on a simplified implementation of Viola-Jones detector [15], enhanced with some heuristics to further reduce the computational burden. For example, the detector does not search for faces in the regions of the image exhibiting low variance.

Camera sensors have limited fields of views and can only perceive a portion of a scene from a single viewpoint. In addition, monocular vision totally lacks 3D information. To mine the entire scene and to deal with occlusions, it is natural to equip visual WSNs with multiview capabilities [16]. Due to bandwidth and efficiency considerations, however, images cannot be routinely shared on the network, so that no dense computation of 3D properties (like disparity maps and depth) can be made. Nevertheless, by codifying geometrical entities, it is still possible to exchange information about regions or points of interest and to contextualize these data in the 3D space. To this end, a coordinator node, knowing the calibration parameters of the camera nodes in the network, may be introduced, so as to translate events from the image coordinates to physical world coordinates. Such approach may produce more robust results as well as a richer description of the scene.

3 Image Mining in WSN

In a distributed reactive application, the mission of the visual WSN is to report any relevant change in the scene, where some real-time constraints should be satisfied. To this

end, one should take into account both local computation and network communication issues to bound the maximum latency in the reaction.

Considering these constraints, but in the need for an efficient mining of the visual scene under examination, we focused on two different approaches: the first one based on simpler detection methods, but more efficient from a computational and storage points of view, the latter giving better performances for more specific object identification tasks while being computationally more intensive.

Moreover, considering a more global analysis and mining of the scene viewed from several different nodes (i.e. multiview), a final processing level for the aggregation of different single in-node data, and the final result of the scene mining needs to be implemented.

In the following we will analyze the above mentioned specific methods for the detection of changes in a scene, in particular regarding the two categories: simple, but quick and efficient change detection algorithms in Section 3.1, and on the other hand, more complex and demanding algorithms for an object detection and identification in Section 3.2. Finally in Section 3.3 we will present methods for the aggregation of hierarchical identifications, made at different level of complexity and from multiple acquiring points of view, in order to be able to correctly mine the set of images to detect events.

3.1 Event and Change Detection

The task of identifying changes, and in particular regions in the image under analysis, taken at different times is a widely diffused topic covering very different disciplines [12]. In this context, the goal is the identification of one or more regions of pixels, relative to a change in the scene, and occurring within a sensible and a priori known area.

In order to perform fast, efficient, and enough reliable methods for this change detection, the generally adopted low-level methods are based on both frame differencing, and statistical methods. This assumption can be either the quick and immediate answer to a simple problems, or a preliminary synthesis for a deeper and more effective higher level analysis. The general model used is mainly based on background subtraction, where a single frame, acquired in a controlled situation, is stored (*BG-image*) and thereafter used as the zero-level, for different types of comparison in the actual processing.

The first methods are very low-level ones and are thought in such a way that it can be possible and feasible to implement them also at firmware level, so to make them real-time operative in an automatic way, examples of these can be:

- Thresholding toward the BG-image
- Contrast modification with respect to the BG-image
- Basic edge detection on the image resulting from the subtraction with the BG-image

Another class of methods is based on statistical histogram analysis. Template histograms are stored of the regions of interests in the BG-image. Then a standard distance, such as the Kullback-Leibler divergence is computed between the region of BG-image, and the region of the actual image. If such distance overcomes a fixed threshold, then a change event is detected.

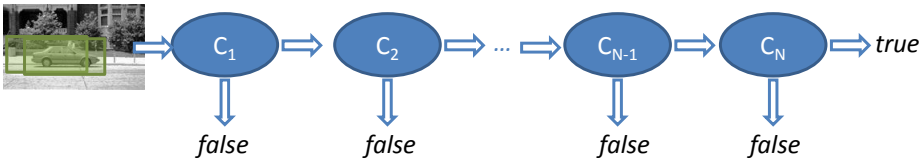

Fig. 1. Cascade of classifier for object detection in a case study of car detection.

3.2 Object Detection

As it is well-known, the problem of detecting classes of objects in cluttered images is challenging. Supervised learning strategies have demonstrated to provide a solution in a wide variety of real cases, but there is still a strong research activity in this field. In the context of visual WSN, the preliminary learning phase may be accomplished off-site, while only the already trained detectors needs to be ported to the network nodes.

Among machine learning methods, a common and efficient one is based on the sliding windows approach; namely rectangular subwindows of the image are tested sequentially, by applying a binary classifier able to distinguish whether they contain an instance of the object class or not. A priori knowledge about the scene or – if available – information already gathered by other nodes in the network may be employed to reduce the search space either by a) disregarding some region in the image and b) looking for rectangular regions within a certain scale range (e.g. rectangular regions covering less than 30% of the whole image area).

For what regards the binary classifiers itself, among various possibilities, the Viola-Jones method is particularly appealing. Indeed, such classifier is based on the use of the so-called Haar-like features, a class of features with limited flexibility but which is known to support effective learning. A Haar-like feature is essentially given by the difference between sum of pixels in rectangular blocks; as such a Haar-like feature is computable in constant time, once the integral image has been computed (see [15] for details). Thus, having enough memory to store the integral image, the feature extraction process needs only limited computational power. Classification is then performed by applying a cascade of classifiers, as shown in Figure 1. A candidate subwindow which fails to meet the acceptance criterion in some stage of the cascade is immediately rejected and no further processed. In this way, only *detection* should go through the entire cascade. In addition, one may train the cascade – for example using *gentle AdaBoost* – in such a way that most of the candidate subwindows that do not correspond to instances of the object class are rejected in the first stages of the cascade, with great computational advantages.

The use of a cascade of classifiers permits also to adapt the response of the detector to the particular use of its output in the network, also in a dynamical fashion, in order to properly react to changes in the internal and external conditions. First of all, the trade-off between reliability of detections and needed computational time may be controlled by adaptive real-time requirements of the overall network. Indeed, the detector may be interrupted at an earlier stage in the cascade, thus producing a quick even though less reliable output, which may be anyhow sufficient for solving the current decision making problem. In the same way, by controlling the threshold in the last stage of the cascade,

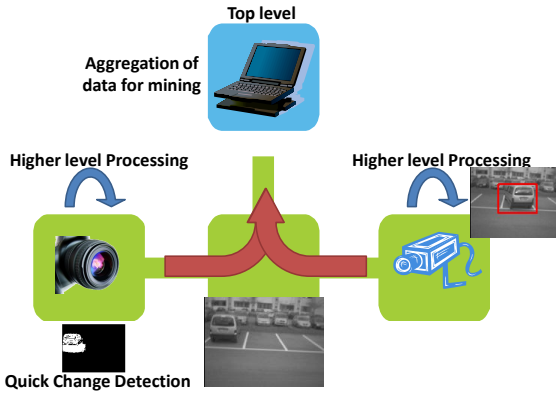

Fig. 2. Architecture of the hierarchical levels of detection, in a car detection case study.

the visual WSN may dynamically select the optimal trade-off between false alarm rate and detection rate needed in a particular context.

3.3 Aggregation of Partial Results

Considering a visual WSN with multiple views of the same scene for image mining purposes, the previously described methods can be managed in a hierarchical fashion, with faster algorithms yielding a primary response to a simple change detection task, and sending their results up through the hierarchy to more performing algorithms, in case of a positive detection. These higher level algorithms, due to their greater computational requirements, are implemented in order to be called only when positive feedback is given by the lower level ones. At the highest level of the hierarchy, a final *decision-maker* algorithm, operates having a not so strict computational and storage constraints, that can aggregate both the results of lower levels processing, but also to combine the multiple views processing, each one operating with the same hierarchy (see Fig. 2).

A common hypothesis for the multiple views aggregation process works on the basis of a given knowledge base of information, relative to the approximate possible positions, or Regions of Interest, where the objects to be mined can be detected. For example the decision maker algorithm can analyze the specific outcomes from different lower-level processing in each single node, and then be able to give a final output using its knowledge and working on the basis of a weighted computation of the single in-node results.

4 Case Study: A Pervasive Infrastructure for Urban Mobility

The Tuscany regional project IPERMOB – " A Pervasive and Heterogeneous Infrastructure to control Urban Mobility in Real-Time" [17] is exploiting an approach based on visual WSNs for building an intelligent system capable of analyzing infomobility data and of providing effective decision support to final users, e.g. citizens, parking managers and municipalities. The partners of the IPERMOB project are a balanced mix

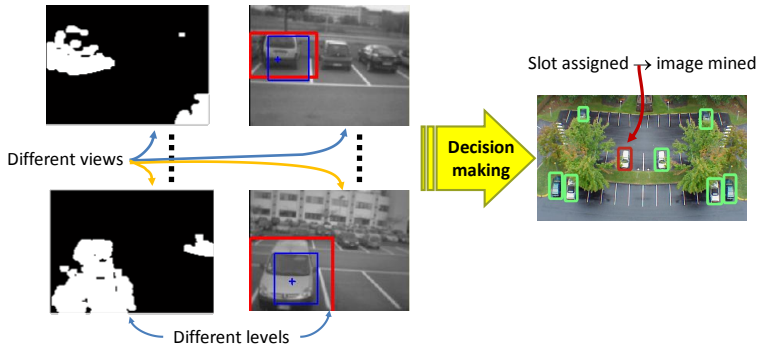

Fig. 3. The aggregation and decision-making process for WSN image mining in IPERMOB Project.

of academical and industrial partners; the work presented in this paper is mainly implemented by the Signals and Images Lab of the Institute of Information Science and Technologies of the National Research Council, and the Real Time Systems Lab of the Scuola Superiore Sant'Anna of Pisa. The IPERMOB project is developing an open, inter-operable system featuring leading-edge technologies for distributed imaging and wireless telecommunication. Supported by the real-world data gathered by pervasive monitoring, IPERMOB aims to become a laboratory for the study of novel models for mobility and to represent a benchmark for technological integration. IPERMOB infrastructure is organized in 3 tiers, corresponding to i) data acquisition and processing, ii) Metropolitan Information Bus (MIB) and iii) end-users applications. The data acquisition and processing tier (by using visual WSNs and Vehicular ad-hoc Networks (VANETs)) is in charge of gathering and processing information, and of transmitting aggregated data to the MIB. The MIB, acting as a middleware application, stores historical data and makes all the information available in a suitable form to the final applications.

In particular, in IPERMOB, visual WSNs are used to monitor parking lots and roads, in order to mine the observed scenes and predicate something about parking lot occupancy and traffic flow. Here, we focus on a parking lot scenario, i.e. in particular a case study at the Pisa International Airport for the landside vehicle flow, in which a set of camera nodes with partially overlapping field of views is in charge of observing and estimating dynamic real-time traffic related information, in particular regarding traffic flow and availability of parking places. It is assumed that each camera knows the geometry of the parking spaces that it is in charge of monitoring. In addition, we assume that a coordinator node – on which the final *decision-maker* algorithms described in Section 3.3 are hosted – knows the full geometry of the parking lot as well as the calibration parameters of the involved cameras, in order to properly aggregate their beliefs.

Figure 3 shows the workflow in the real case study. In a typical scenario, some of the camera nodes in the first level of the hierarchy detects a change in the scene (by using the methods reported in Section 3.1) and triggers object detection methods in the remaining camera nodes that are in charge of monitoring that area. The aggregation and decision making process are finally performed by the coordinator node.

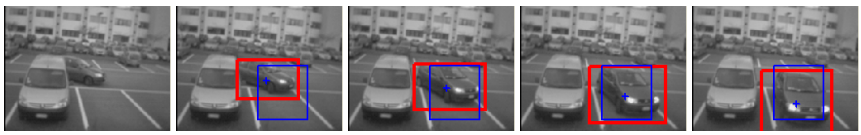

Fig. 4. Car detection and analysis of parking lot occupancy status.

Fig. 5. Detection of vehicles on a road for traffic flow.

For what regards object detection, several cascade of classifiers have been trained, each one specialized in detecting particular view of cars (e.g. front, rear and side views). A large set of labeled acquisition has been made of the real case study and used for training purposes. We notice that first stages in the cascade have low computational complexity but reasonable performance (that is almost 100% detection rate but even 50% false alarm rate). Composition of the stages entails then high performance of the entire cascade, since both overall detection rate and overall false alarm rate are just the product of the detection rate and false alarm rate of the single stages. For example, using $N = 10$ stages in the cascade, each one with detection rate $\nu = 99.5\%$ and false alarm rate $\mu = 50\%$, one gets an overall detection rate $\nu_{\text{global}} = \nu^N \approx 95\%$ and false alarm rate $\mu_{\text{global}} = \mu^N \approx 0.1\%$.

Figures 4 and 5 show examples of two mined image sequences in the framework of the project, respectively for parking availability and traffic flow.

5 Conclusions and Further Work

In this paper, strategies for tackling the image mining problem over visual sensor networks have been presented, taking into account both the structural limits of embedded platforms and the possibilities gained by performing pervasive computing and by the greatly flexible organization of the network topology.

A multi-node processing method has been introduced to decompose a complex computer vision task into a hierarchy of computationally simpler problems to be solved over the nodes of the network. Besides computational advantages, such hierarchical decision making approach appears to be more robust and fault-tolerant.

Such ideas and preliminary results are illustrated by means of an application scenario regarding infomobility, which is currently in development within the IPERMOB project. Set-up and commissioning of the visual sensor network in the project testbed will start in fall 2010. It is expected that the testbed will give precise information whether the presented approach is mature for technological transfer.

Acknowledgements

This work has been partially supported by POR CReO Tuscany Project "IPERMOB" – A Pervasive and Heterogeneous Infrastructure to control Urban Mobility in Real-Time (Regional Decree N. 360, 01/02/2010). We would like to thank Dr.Eng. M. De Pascale (Società Aeroporto Toscano Galileo Galilei S.p.A.) of the Pisa International Airport for his kind support and for providing access to the testing area.

References

1. Soro, S., Heinzelman, W.: A survey of visual sensor networks. Advances in Multimedia (2009) Article ID 640386, 21 pages
2. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans. PAMI 30 (2008) 555–560
3. Colantonio, S. et al.: An intelligent and integrated platform for supporting the management of chronic heart failure patients, Computers in Cardiology (2008) 897 – 900
4. Salvetti, O., Cetin, E.A., Pauwels, E.: Special issue on human-activity analysis in multimedia data. Eurasip Journal on Advances in Signal Processing (2008) article n. 293453
5. Pagano, P., Piga, F., Lipari, G., Liang, Y.: Visual tracking using sensor networks. In: Proc. 2nd Int. Conf. Simulation Tools and Techniques, ICST (2009) 1–10
6. Kundur, D., Lin, C.Y., Lu, C.S.: Visual sensor networks. EURASIP Journal on Advances in Signal Processing Signal Processing 2007 (2007) Article ID 21515, 3 pages
7. Feng, W., Code, B., Kaiser, E.C., Shea, M., chang Feng, W., Bavoil, L.: Panoptes: scalable low-power video sensor networking technologies. In: ACM Multimedia. (2003) 562–571
8. Hengstler, S. et al.: Mesheye: a hybrid-resolution smart camera mote. In: Proc. 6th Int. Conf. Inf. proc. in sensor networks, ACM (2007) 360–369
9. Rowe, A. et al.: CMUcam3: An open programmable embedded vision sensor. Technical Report CMU-RI-TR-07-13, Robotics Institute, Pittsburgh, PA (2007)
10. Chen, P. et al.: Citric: A low-bandwidth wireless camera network platform. In: Distributed Smart Cameras. (2008) 1–10
11. Remagnino, P., Shihab, A.I., Jones, G.A.: Distributed intelligence for multi-camera visual surveillance. Pattern Recognition 37 (2004) 675–689
12. Radke, R.J., Andra, S., Al-Kofahi, O., Roysam, B.: Image change detection algorithms: a systematic survey. IEEE Transactions on Image Processing 14 (2005) 294–307
13. Yan, T., Ganesan, D., Manmatha, R.: Distributed image search in camera sensor networks. In Abdelzaher, T.F., Martonosi, M., Wolisz, A., eds.: SenSys, ACM (2008) 155–168
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60 (2004) 91–110
15. Viola, P.A., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57 (2004) 137–154
16. Pagano, P., Piga, F., Liang, Y.: Real-time multi-view vision systems using WSNs. In: Proc. ACM Symp. Applied Comp., ACM (2009) 2191–2196
17. IPERMOB: A Pervasive and Heterogeneous Infrastructure to control Urban Mobility in Real-Time. http://www.ipermob.org/ (2010) Last retrieved Feb 25, 2010.